



Leibniz-Institut für
Astrophysik Potsdam

Research Environments, Data Publication and Virtual Observatory

**Kristin Riebe,
Jochen Klar,
Sergei Schmalz,
Harry Enke**

AIP Colloquium, 05.12.2014

Outline

- Research Environments and Data processing
- Data Publication
- Virtual Observatory

- Only some aspects of our work are presented
- Supercomputing will be presented additionally soon
- Afternoon hands-on session will focus on data access and SQL

Some Infrastructure Projects

- German Astrophysical Virtual Observatory (GAVO), AIP Co-PI, 2003 – 2016
- AstroGrid-D, AIP PI, 2005-2009
- WissGrid, AIP in Steering Committee, 2009-2011
- DFG Project Radieschen (General Set-up for Interdisciplinary Research Data Infrastructure), AIP in Steering Committee, 2011-2013
- DFG Project Sustainability of Virtual Research Environments, 2013 - 2014

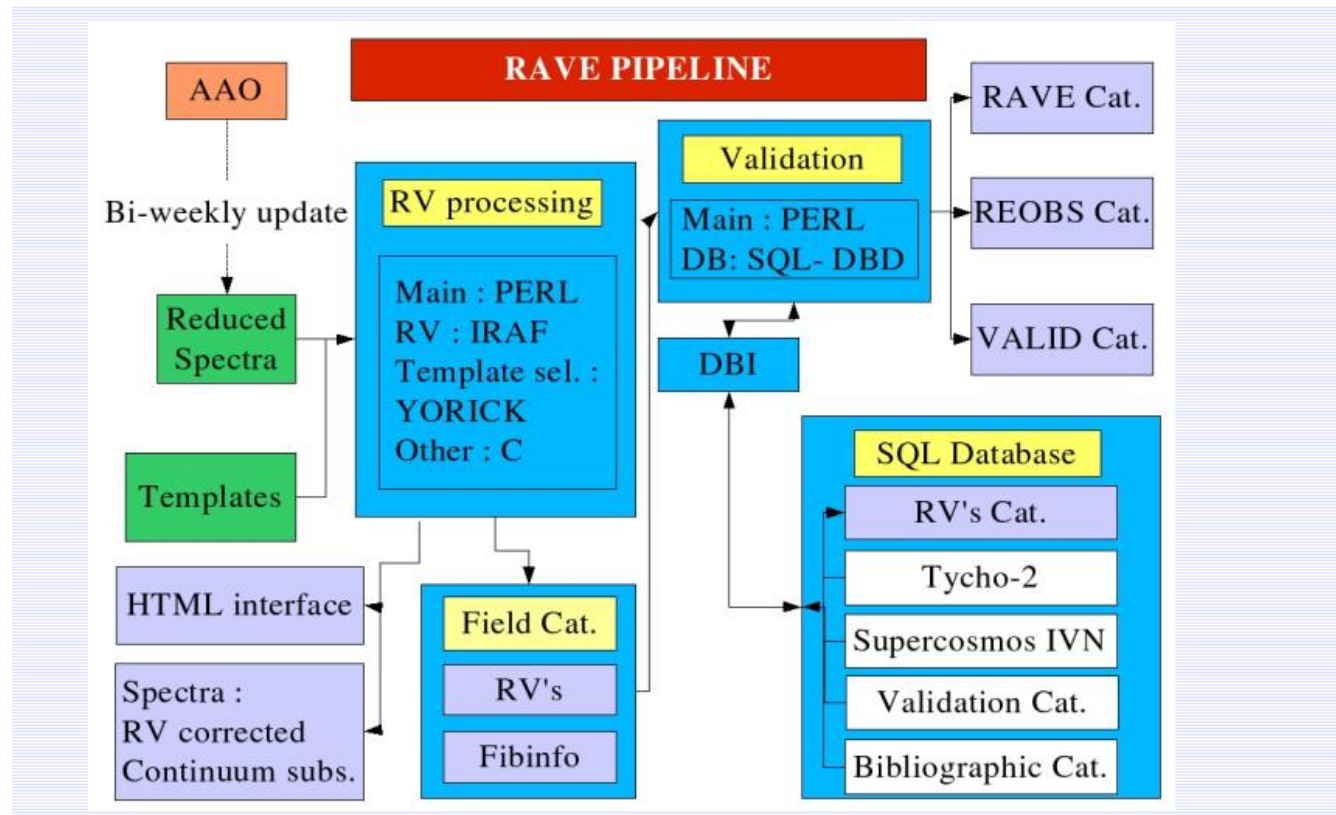


Leibniz-Institut für
Astrophysik Potsdam

Research Environments and Data processing

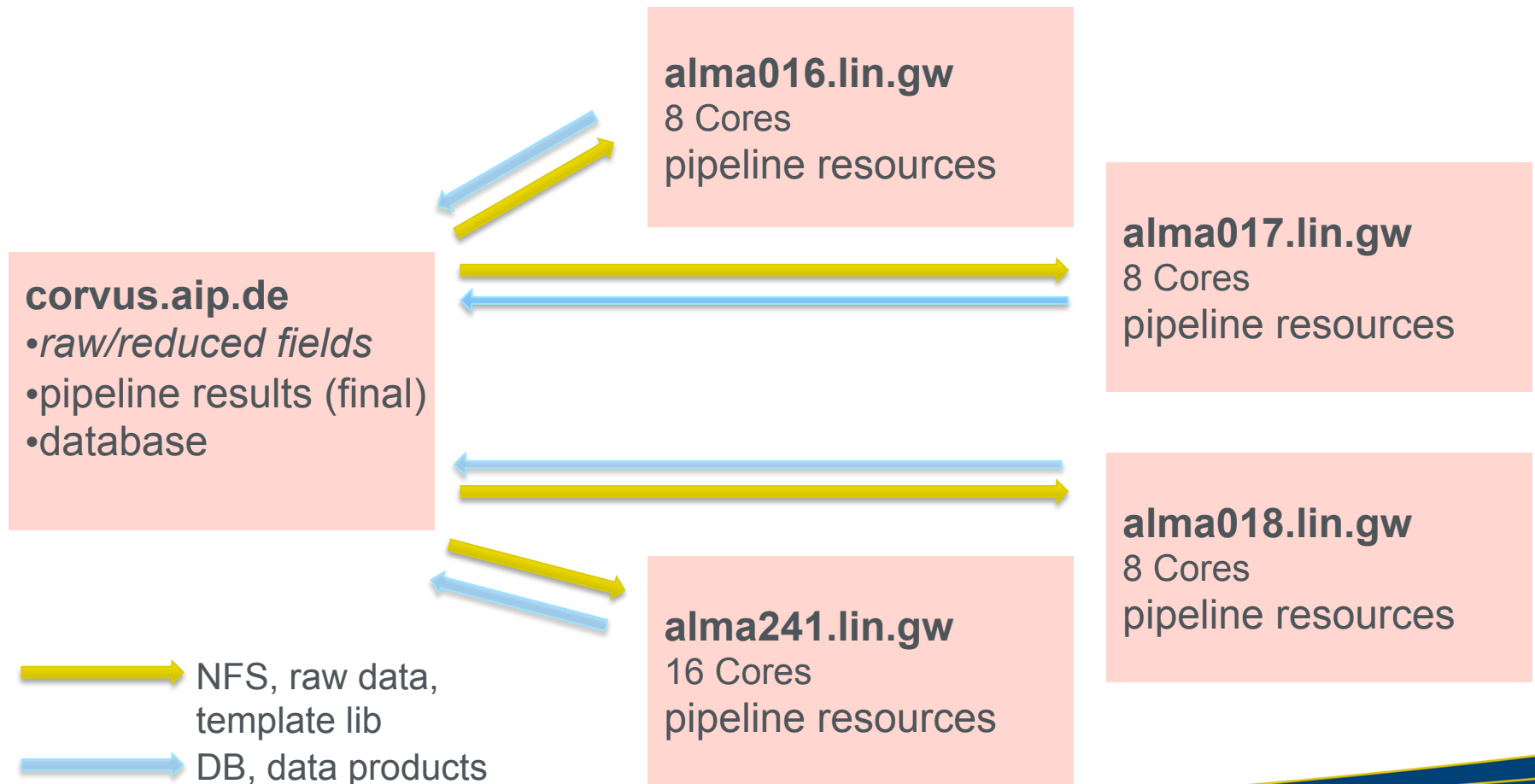
RAVE Pipeline & Archive

2009/2010: Rewrite pipelining process:
from single core pipeline



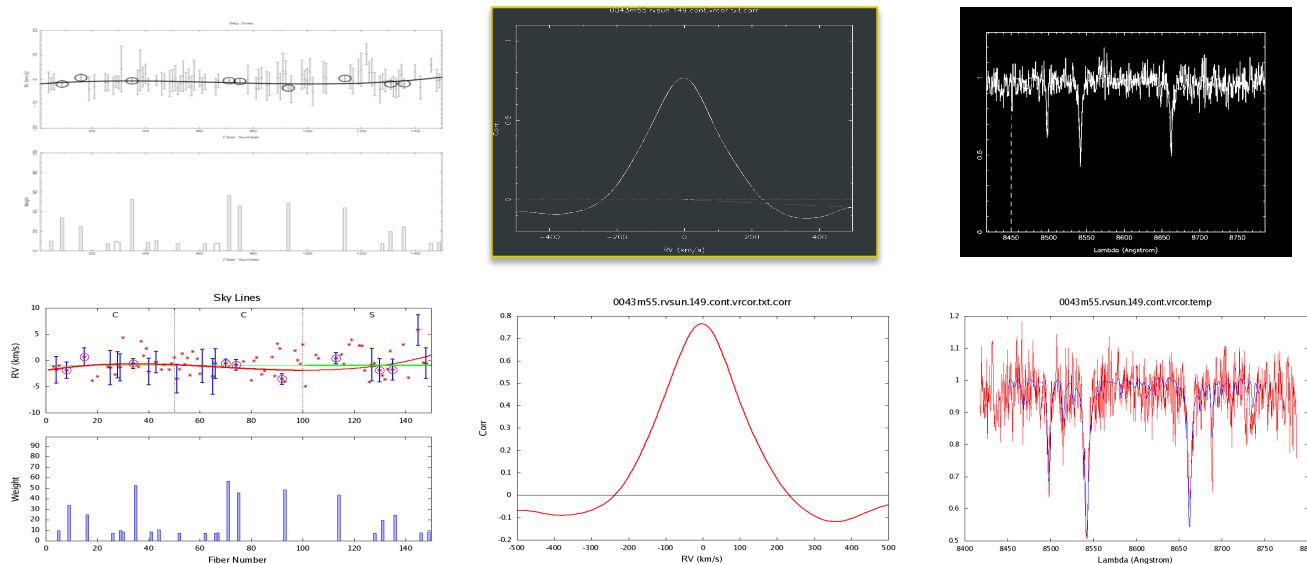
RAVE Pipeline & Archive

2009/2010: Rewrite pipelining process:
to N-Core multi node processing (Speedup $\sim$ N cores)



RAVE Pipeline & Archive

- source control for pipeline (preprocessing+ SPARV)
- tracking of individual fibers instead of whole fields (= 150 fibers)
- Improved data products



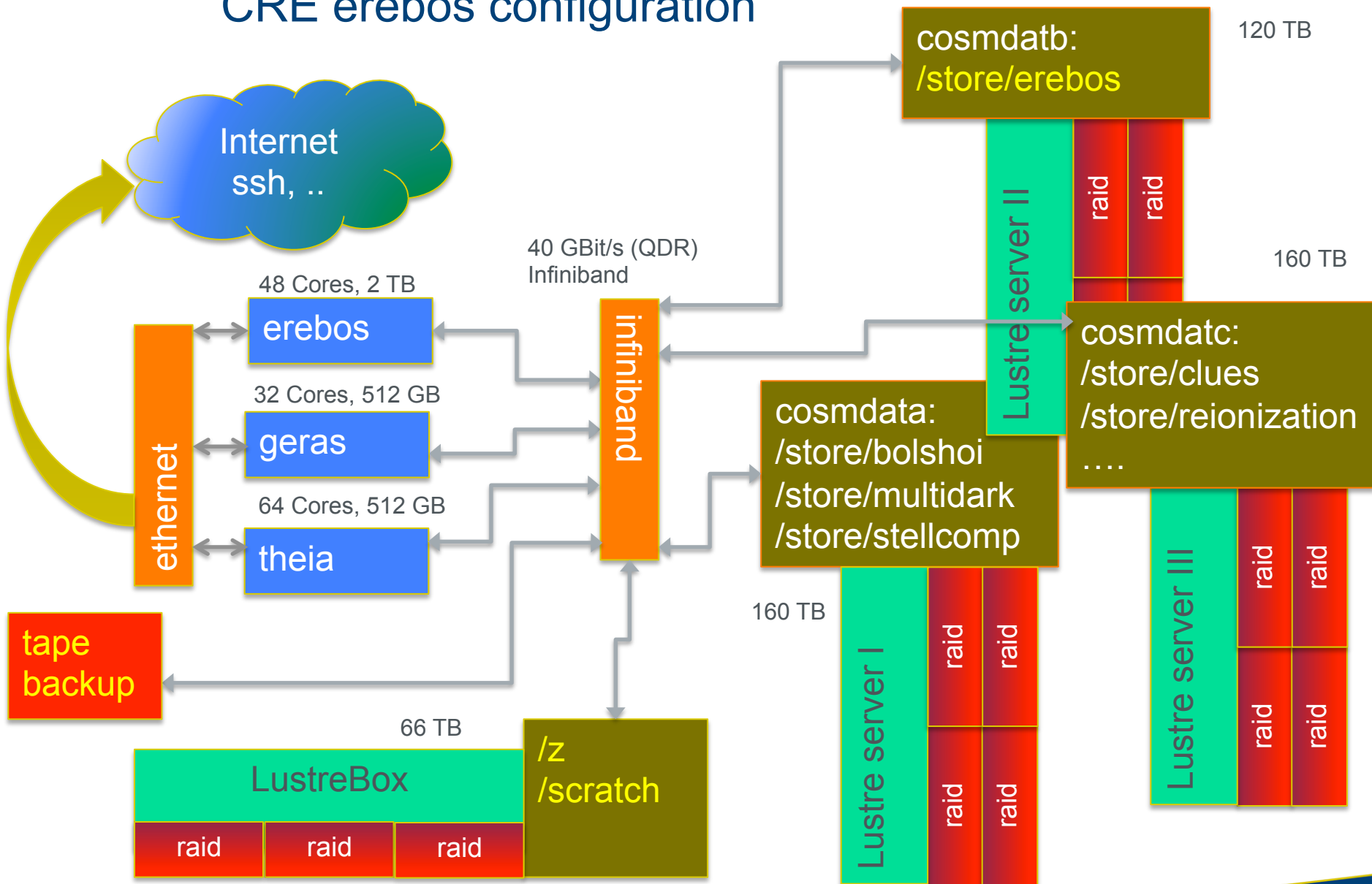
- quality control and observer comments extraction from fits files
- provenance information for each field and fiber
- UCD for each column in database (Unified Content Descriptors)
- cross match with multiple catalogs
- web based interface to all data, differentiated access levels (public, collaboration, DMG), sql queries

Collaborative Research Environment (CRE)

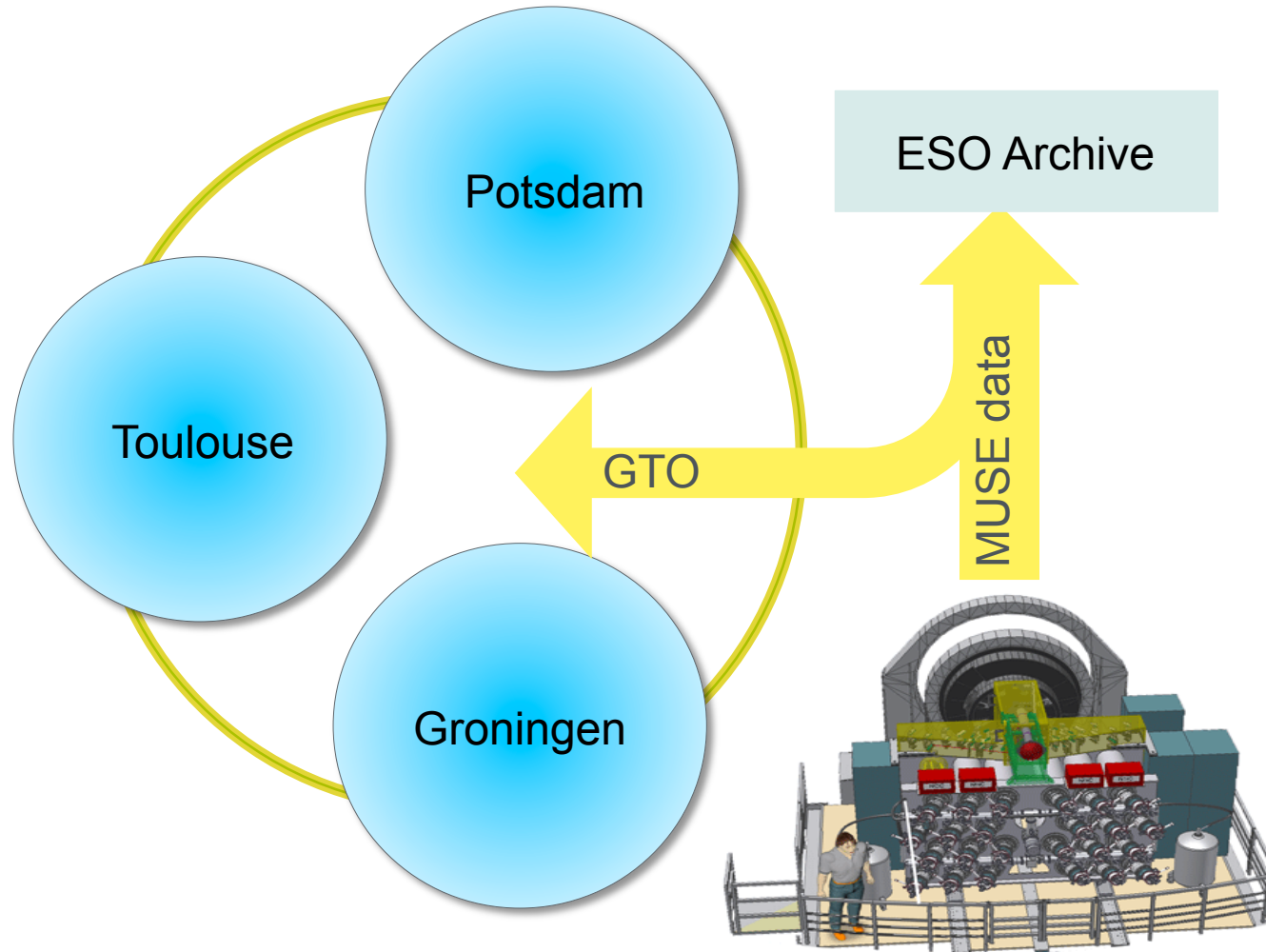
CLUES project as example:

- huge data space
 - for raw data
 - for post processing product data
- post processing tools
- scientific programs, compiler, libs, viewers...
- management procedures for a collaboration
 - user accounts etc. managed by collaboration
 - membership not bound by relation to institute or university
 - mailing list
 - accessible, secure machines
 - project web space for collaboration results

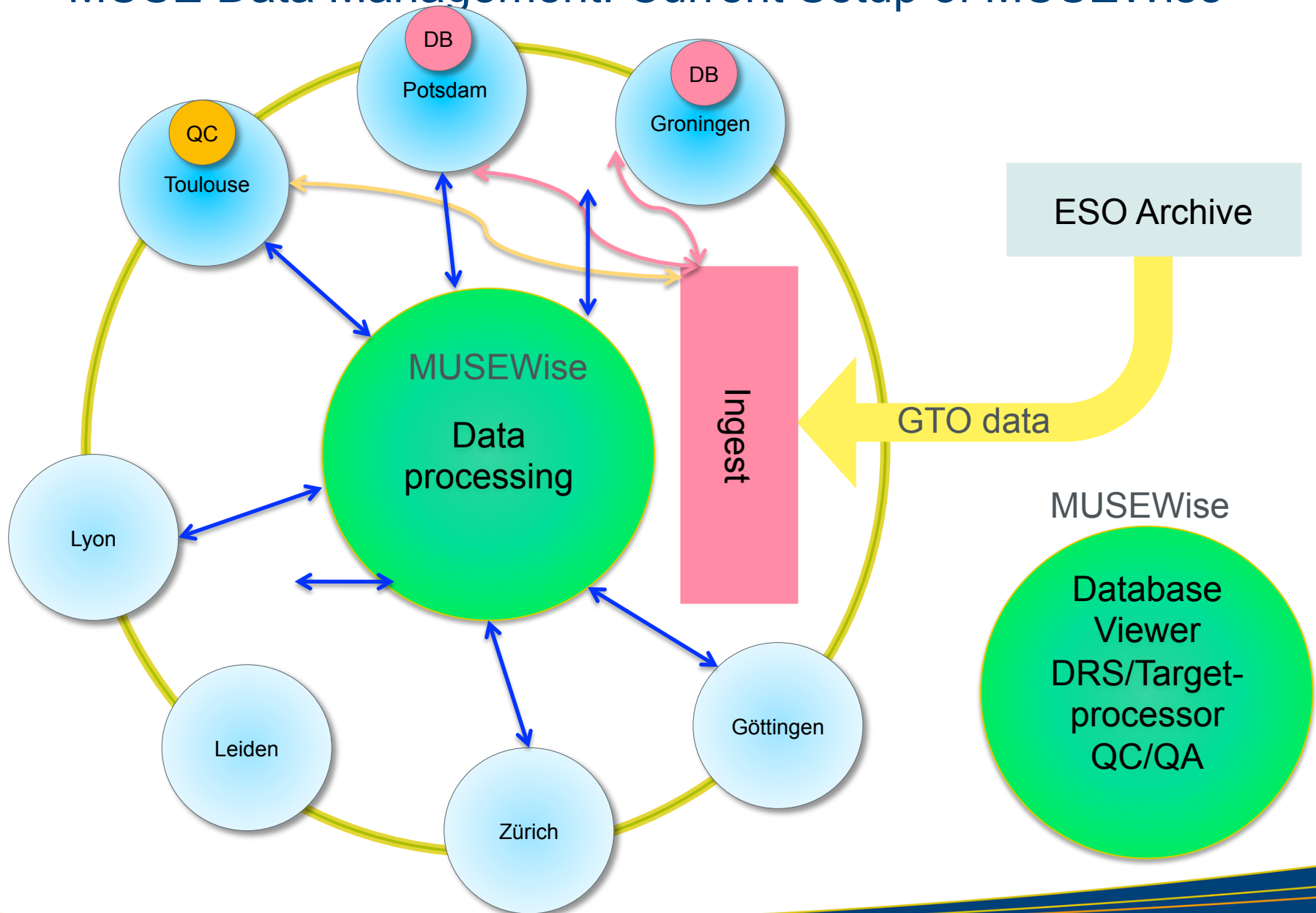
CRE erebos configuration



MUSE Data Management: Raw Data Flow



MUSE Data Management: Current Setup of MUSEWise



Digitization of astronomical photographic plates

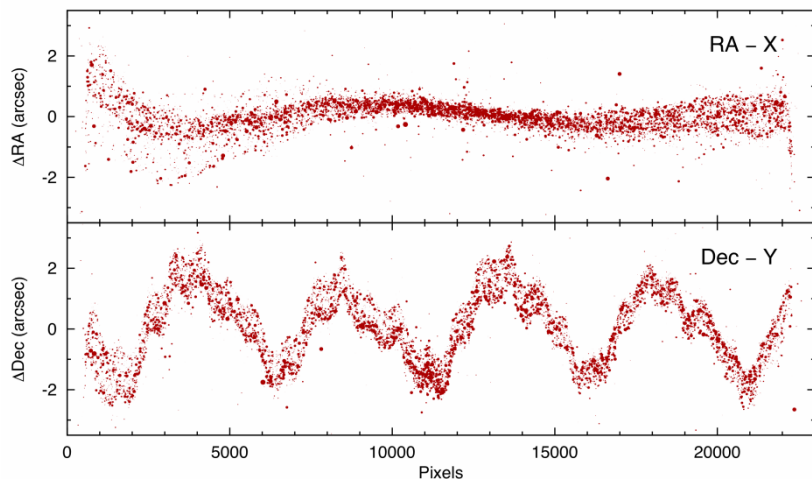
DFG project of AIP, Dr. Remeis Sternwarte Bamberg and Sternwarte Hamburg

(also continuation of former efforts (WFPD, M.+K. Tsvetkov))

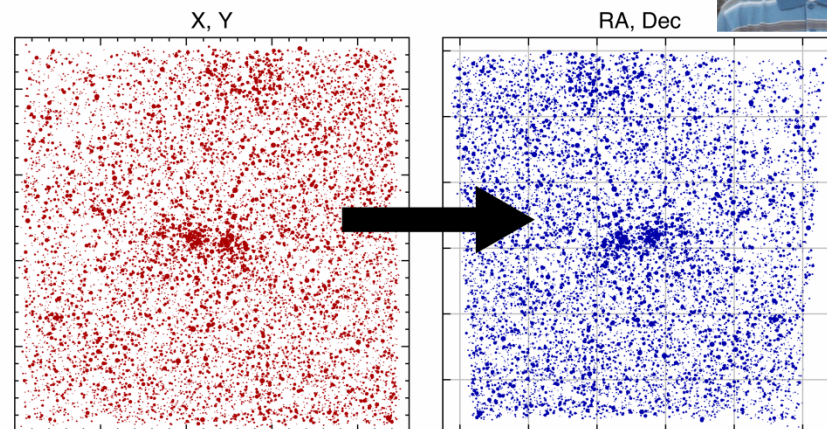
Digitizing plates, extraction of logbook metadata

- 10000 AIP
- 40000 Bamberg
- 35000 Hamburg
- Definition of FITS header extensions for digitized plates
- Source catalog extraction from images
- Photometric information with suitable calibration for many objects (by crossmatch with other catalogs)
- pyPlate software package and pipeline
- Working on calibration of spectral plates ongoing
- Archive and publication via webinterface @ AIP
- APPLAUSE (Archive of Photographic PLates for Astronomical USE)
(<https://www.plate-archive.org>)

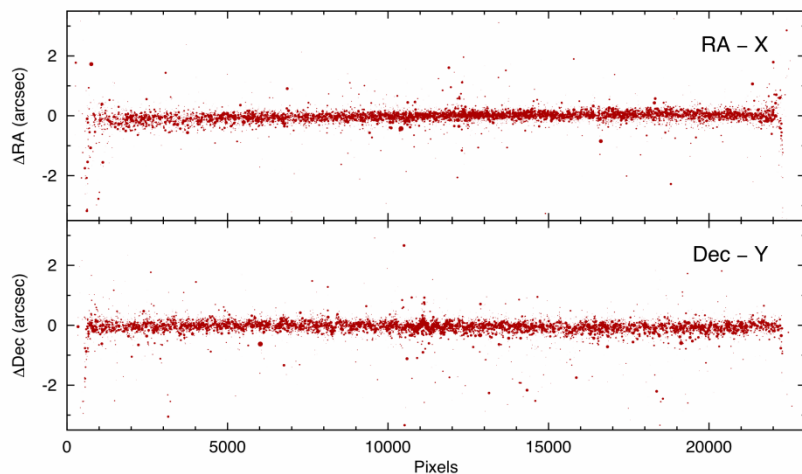
Digitization of astronomical photographic plates



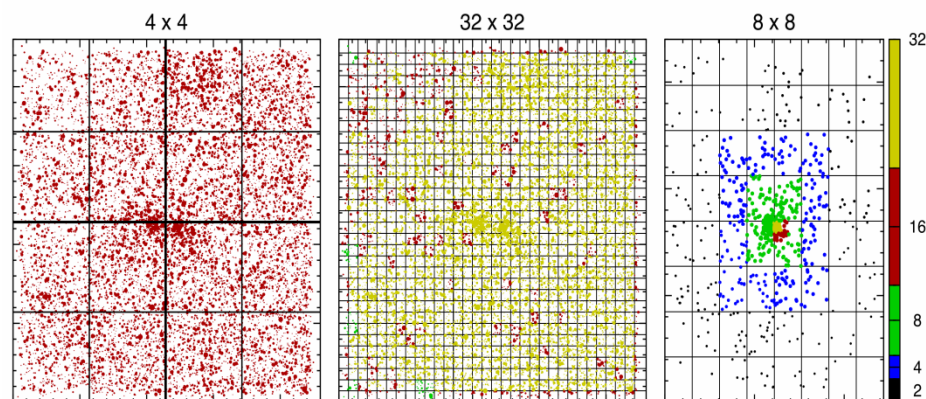
Initial solution with Astrometry.net



Initial solution with Astrometry.net software



Improved solution with SCAMP



Improving solution with SCAMP software in sub-fields

more: https://escience.aip.de/ag2014/Tuvikene_AG2014.pdf

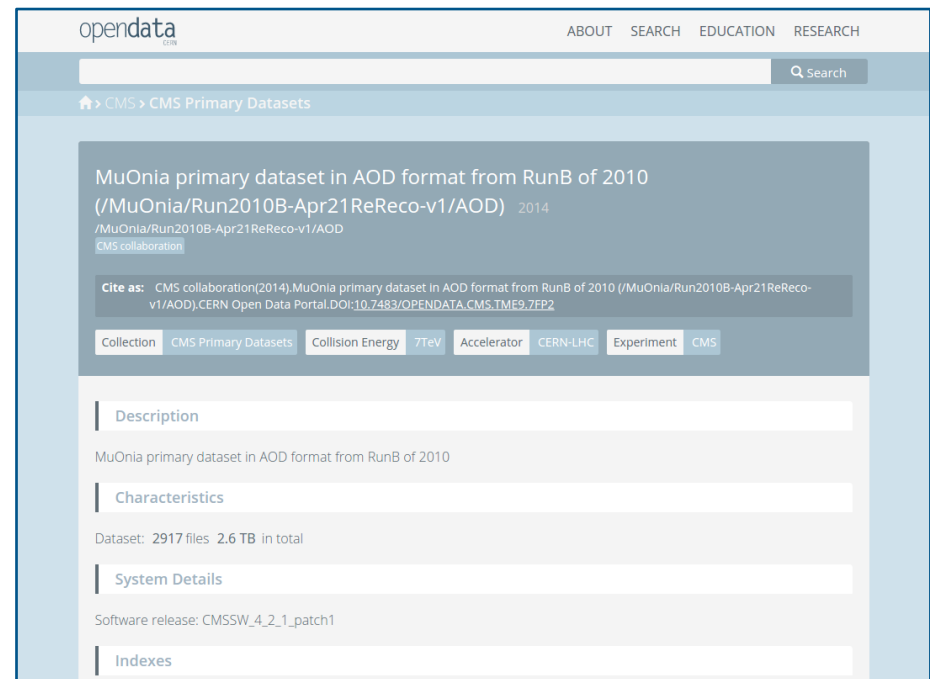


Leibniz-Institut für
Astrophysik Potsdam

Database Publication

Data publication

- Publication of research data is pushed forward by:
 - Politics: European Commission, Bundesregierung
 - Funders: DFG, BMBF, Horizon 2020
 - Journals: Nature, PLOS
 - increasing amount of Scientists
- Recent example: CERN
 - opendata.cern.ch
- Most active:
 - Geosciences
 - Climatology
 - Humanities
 - Social Sciences
 - Astronomy !!!



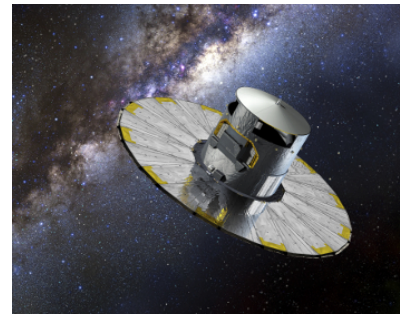
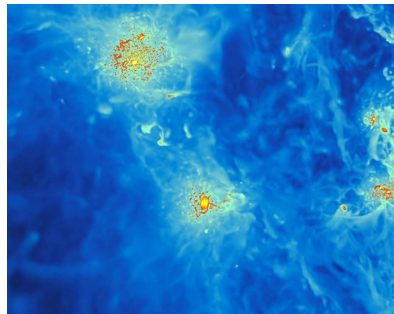
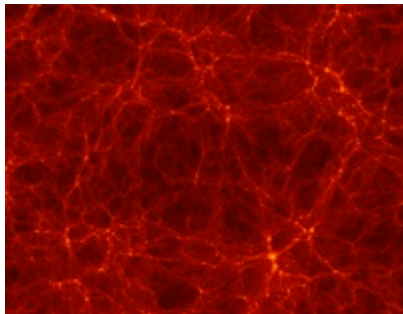
Astronomy on the forefront



SQL (or ADQL) is the de-facto standard in astronomy.

Some issues ...

- Custom made applications, hardly reusable.
- Some Closed-Source components.
- Considerable maintenance *per* website.
- More and more projects want to or have to publish research data in a similar way.
- Performance issues for very large datasets.



What do we need?

- Single software stack to publish SQL databases.
- Highly configurable, but still maintainable.
- Suitable performance for huge datasets.
- State-the-art technology, modern user interface.
- VO compliant. Fully Open-Source.

Our Solution

Daiquiri: A web framework for scientific database publication.

PaQu & Spider Engine: Parallel SQL database queries.

Daiquiri

- Highly customizable framework for database publication
- Many applications – one code base
- Queued SQL queries including SQL query assistance
- Can be used together with PaQu for Parallel queries
- Query result table viewer, quick plotting tool
- User database space, download in different formats
- VO compatible meta data management for databases
- UWS (VO Universal Worker Service) interface
- SAMP connection to VO tools (e.g. Topcat)
- User management, Contact messages, WordPress integration, Meeting organization, ...



Daiquiri

- PHP5, Zend Framework
- MySQL database + some custom MySQL plugin
- Gearman application framework
- WordPress
- jQuery, Bootstrap, Angular JS, CodeMirror for the client
- Separation of
 - The web application (*app*)
 - Core functionality (*daiquiri*)
- Licensed under AGPL v3
- published on github
 - <https://github.com/jochenklar/daiquiri>
 - <https://github.com/jochenklar/daiquiri-app>



PaQu & Spider Engine

Spider Engine

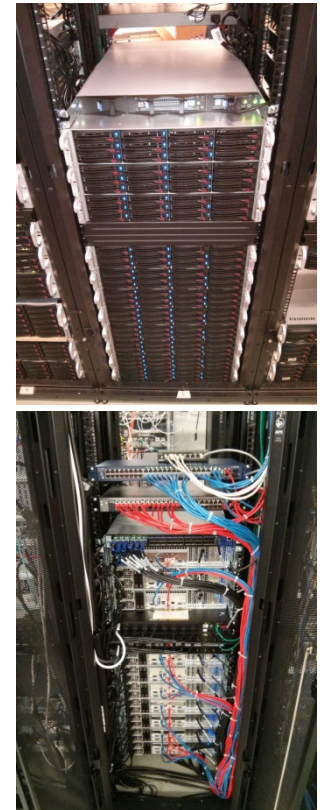
- distribution of the database over 10 nodes
- orchestrated by a head node
- simple queries can be performed in parallel

PaQu

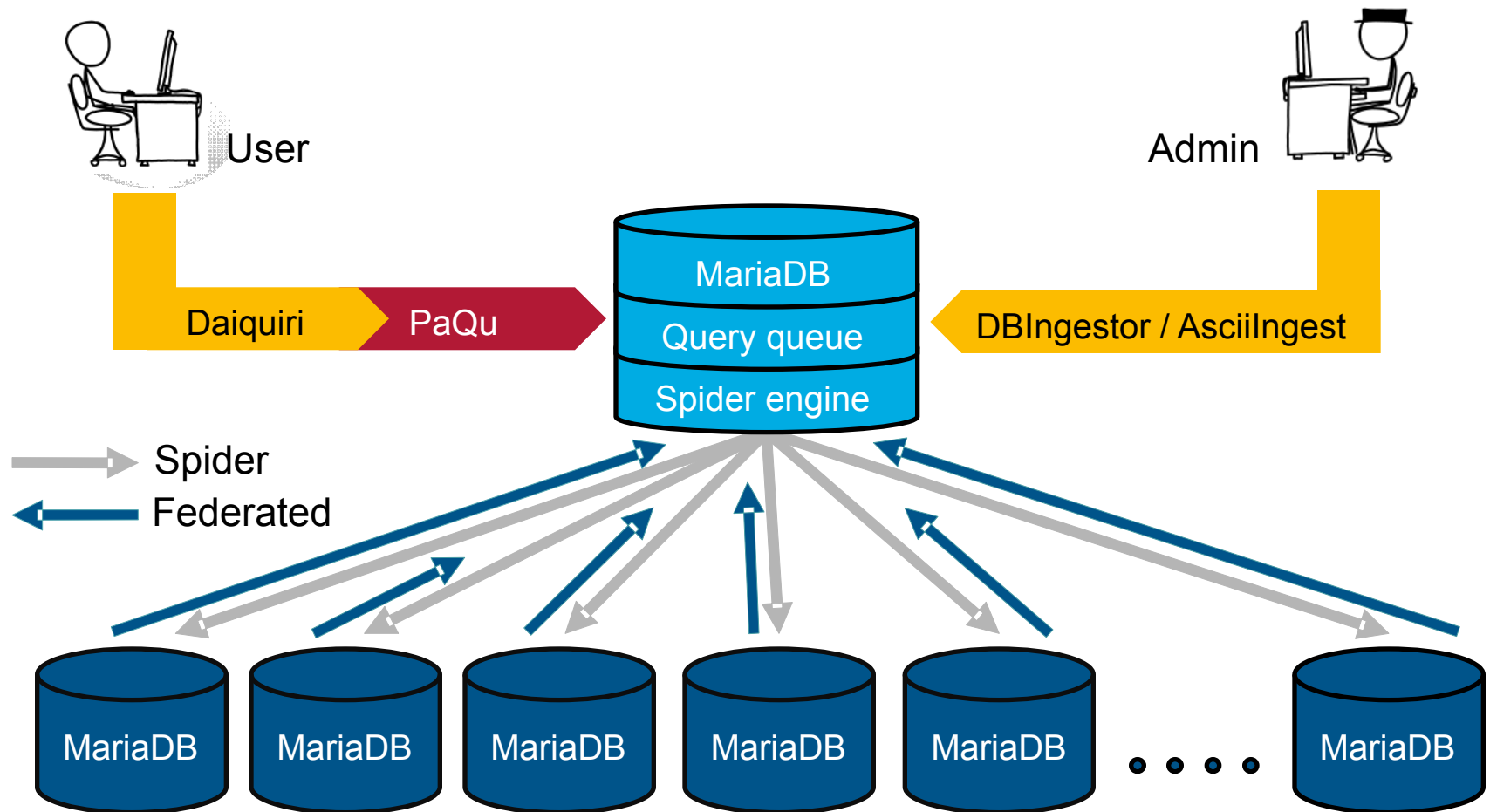
- reformulates queries for joins, aggregates, etc.
- e.g. count rows
 - count on each node
 - sum on head node
- head node collects data via federated tables



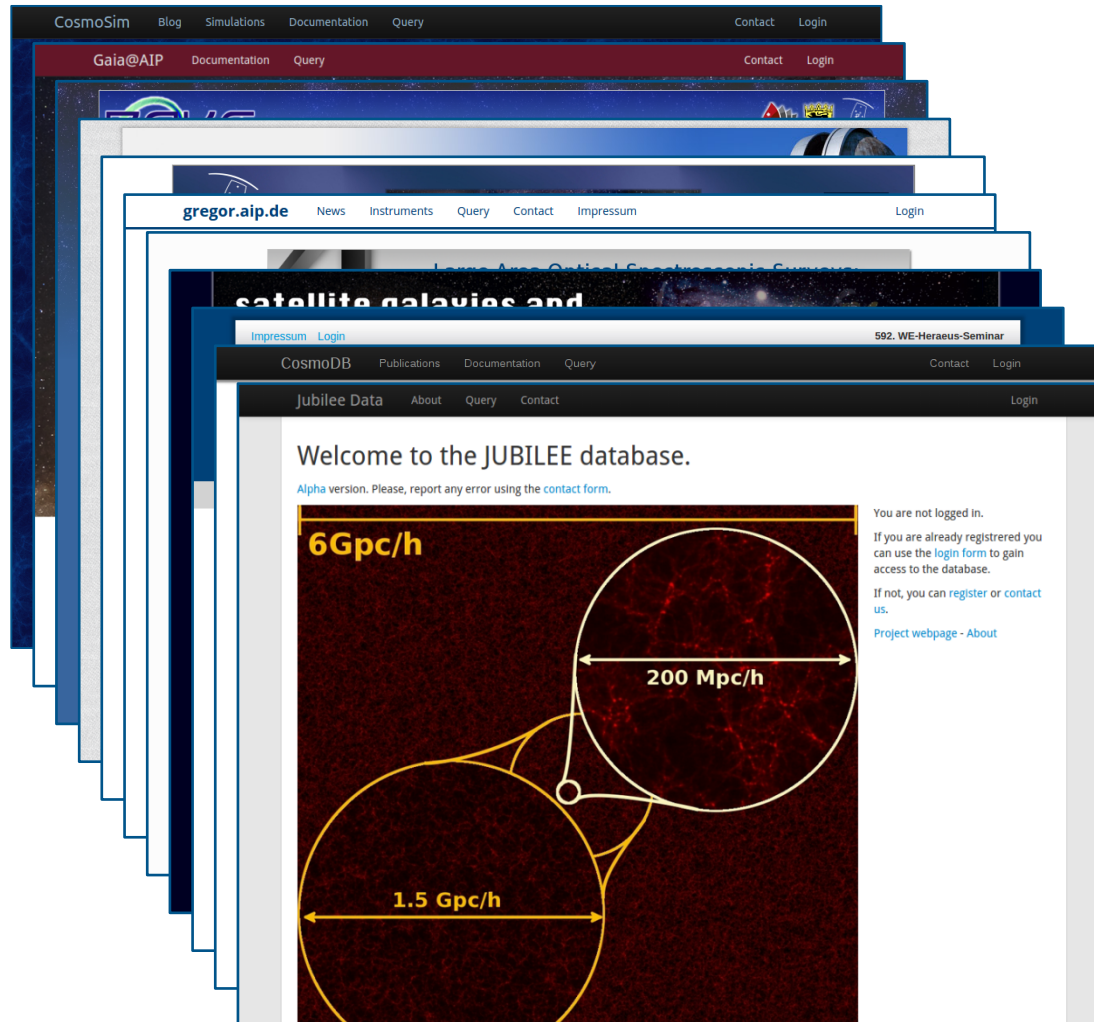
Adrian Partl



The full picture



What is it used for?



- CosmoSim database
- Gaia database
- RAVE database
- APPLAUSE plate archive
- CLUES user management
- GREGOR website
- 4MOST workshop 2012
- Dwarfs workshop 2014
- Heraeus-Seminar 2015
- CosmoDB @ Tartu
- JUBILEE @ Madrid



Leibniz-Institut für
Astrophysik Potsdam

Virtual Observatory

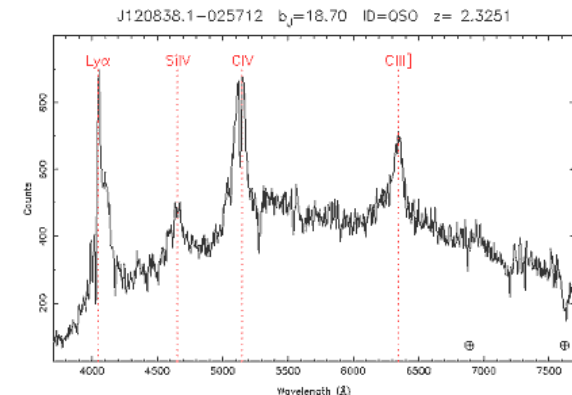
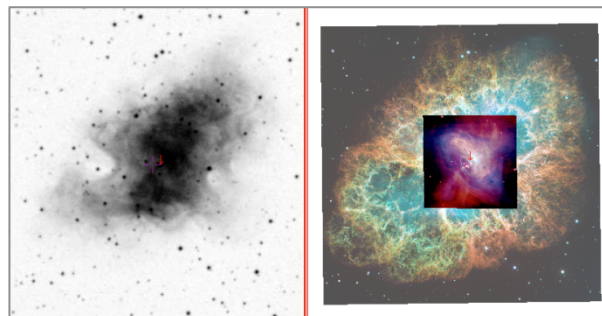
Motivation

Variety of data

- Multi-wavelength data
- Different types of data:
 - Photometry, astrometry, spectra, time series, simulations
- Different people, institutions and telescopes
- Complex, individual data formats



How to handle this?



International Virtual Observatory Alliance



Participating member organizations

Goals of the VO



- Data access
 - Allow and improve access to all kinds of data
- Data discovery
 - Make it easy to find data for given object/position/type/...
- Data publication and archives
 - Publish and archive data
 - Useful metadata
- Software
 - To discover, access, handle and exchange data

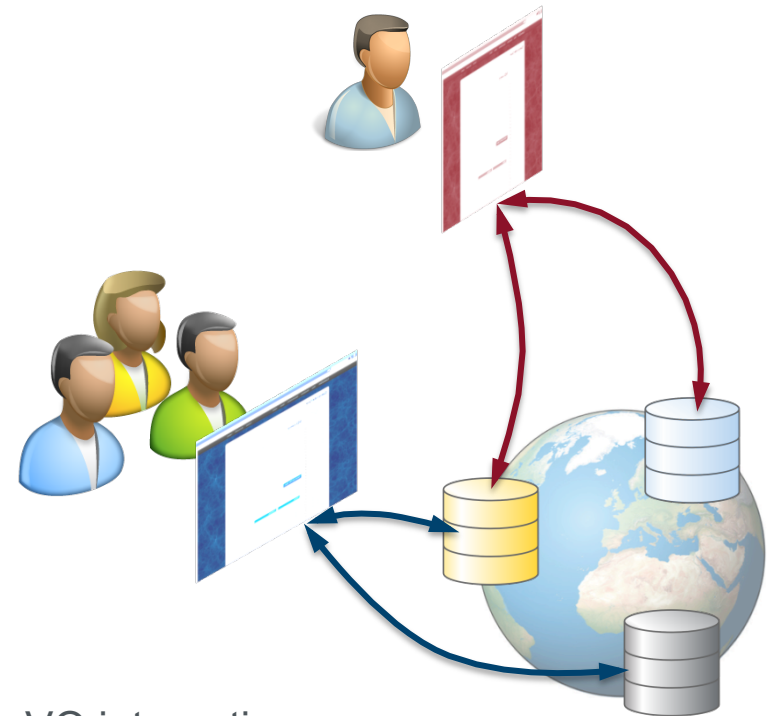
How?

Standards and
interoperability!

=> Maximize scientific output of existing data

Virtual Observatory

- Standards
 - Data formats (e.g. VO-table)
 - Data descriptions (UCDs, provenance)
 - Access (tables, images, spectra)
 - Exchange (SAMP)
- Interoperable tools & applications
- Distributed data centers
 - All over the world, using the same standards for VO integration
- VO registry
 - „yellow pages“ of the VO services and data
 - VO services and their data collections registered
 - easy data discovery



GAVO

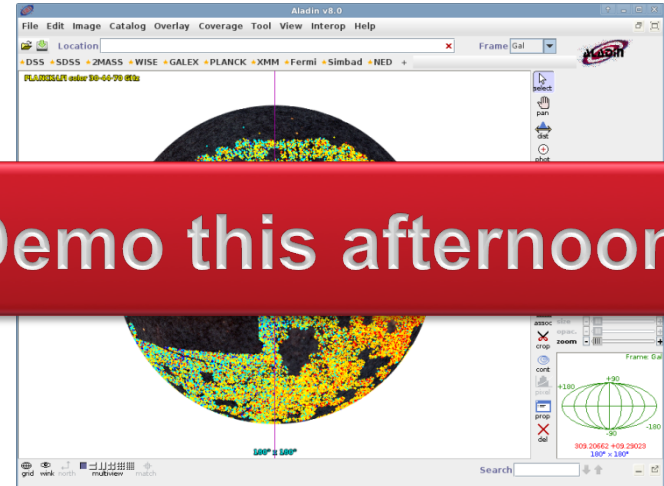
- German Astrophysical Virtual Observatory
- German contribution to the IVOA
- People from Heidelberg, Potsdam, Garching, Munich, Tübingen, Bonn
- Publication of
 - Observational databases
 - Simulation databases
- GAVO data center software
- Standard creation and implementation
- Outreach:
 - VO Days, AG-meetings, Tutorials



VO at AIP

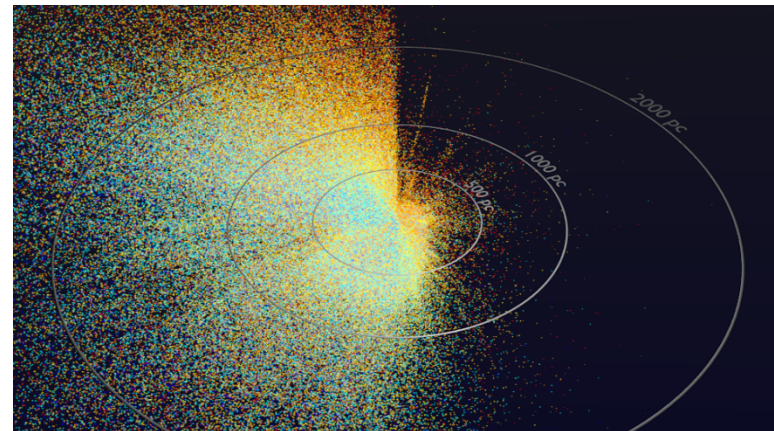
Observational databases

- RAVE
 - Pipeline
 - Database structure
- Digital plates archive
- SDSS mirror



Demo this afternoon!

RAVE data in Aladin



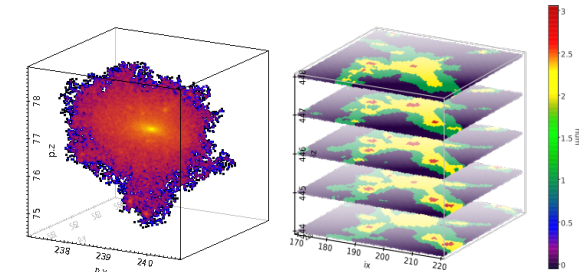
RAVE stars in Blender

VO at AIP

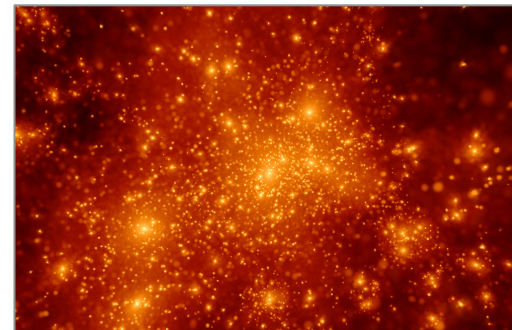
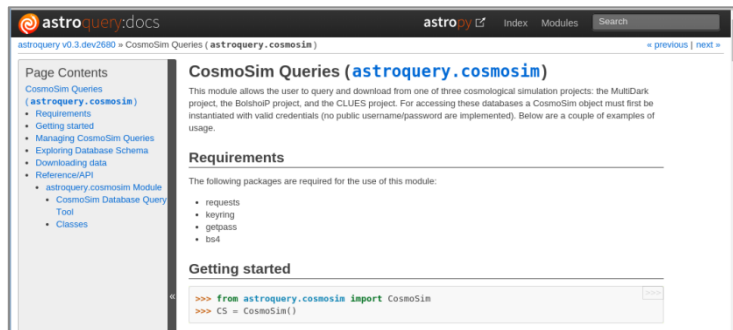
Simulation databases

- CosmoSim
 - Database with cosmological simulations
 - 250 registered users from 35 countries
 - MultiDark and CLUES project, 6 sim., 64 – 1000 Mpc/h
 - Data curation, data structures and documentation
 - Use case for Daiquiri+Spider+PaQu
 - Facilitating the work flow (DBIngestor + extensions)
 - astroquery.cosmosim package (Austen Groener) for scripted access

Demo this afternoon!



CosmoSim data in Topcat

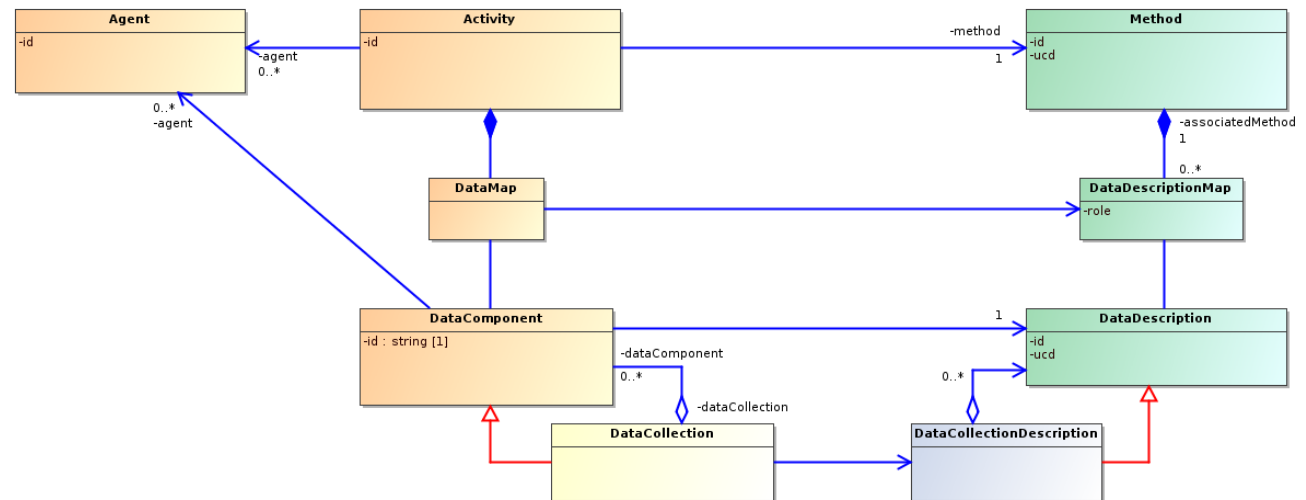


*Visualisation
with PMViewer*

VO at AIP

Standards implementation & development

- UWS for job handling
- SAMP for data exchange
- Drafts for Provenance Data Model
 - Describe origin of data, workflow
 - Retrace bugs
 - Attribution



VO at AIP

Tools

- UWS client

- <http://github.com/adrpar/uws-client>
- Python client to access any UWS-enabled web service
- Submit queries, download results, delete jobs

```
uws --host www.cosmosim.org/uws/query --user xxx --password xxx \  
    job new \  
    query="SELECT x,y,z FROM MDR1.FOF LIMIT 10" \  
    --run
```

```
uws --host www.rave-survey.org/uws/query --user xxx --password xxx \  
    job new \  
    query="SELECT RAdeg, DEdeg FROM RAVEPUB_DR4.RAVE_DR4 LIMIT 10" \  
    --run
```

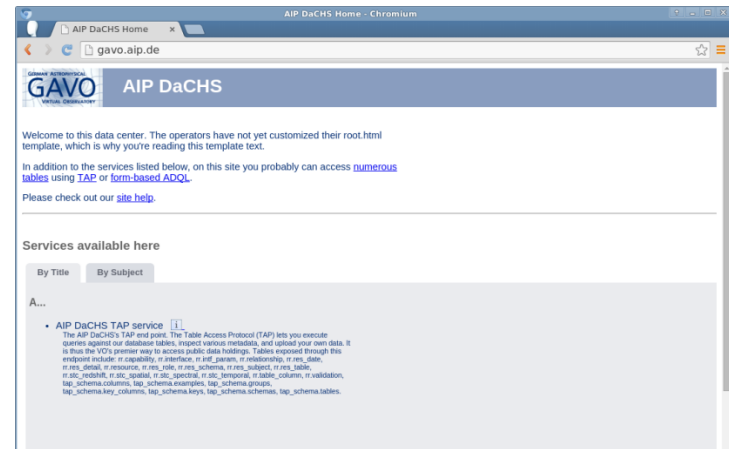
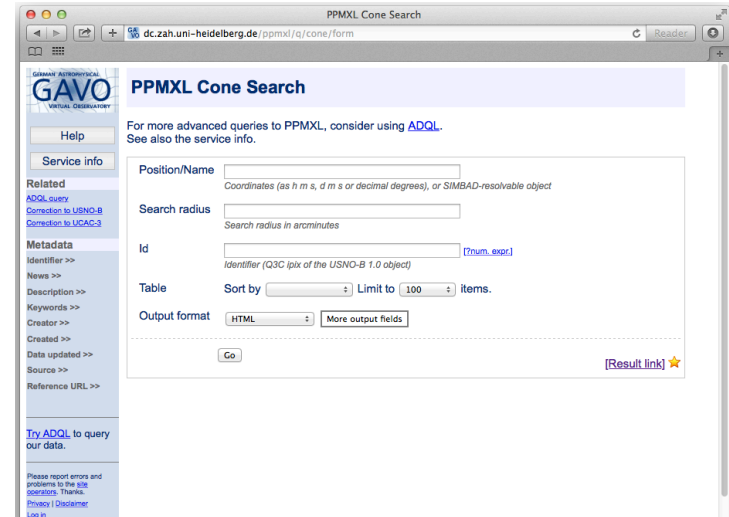
- mysqldump-vo

- <http://github.com/adrpar/mysqldump-vo>
- Directly export data from a database as VO table

VO at AIP

GAVO data center

- VO conform publication of data
 - includes TAP with ADQL, SIA, SSA, ...
- Supports custom forms
 - e.g. Cone search
- Registry included
 - data immediately visible
 - accessible via VO tools like Topcat



Summary

- E-Science since 2003
- Pipeline and data management for RAVE
- Implementation of CREs for CLUES, MUSE, etc.
- Development of Daiquiri web framework for data publications
- High performance databases with PaQu and Spider engine
- Data curation, VO conform publication
- VO standards and interoperability

Tutorials – afternoon, 13:30 pm

- First steps with Aladin (20 min)
- RAVE with Topcat and Aladin (20 min)
- CosmoSim + Topcat (20 min)
- AIP's Data Access Service (20 min)

- Open questions (10 min)