IST-5-033437

The Chemomentum Data Services

A flexible solution for data handling in UNICORE

Katharina Rasch, Robert Schöne, Hartmut Mix - Technische Universität Dresden, ZIH
Vitaliy Ostropytskyy, Werner Dubitzky – University of Dublin
Mathilde Romberg – Forschungszentrum Jülich

Outline

- Chemomentum project overview
- Data management features
- Technical details
- User client

Chemomentum project overview

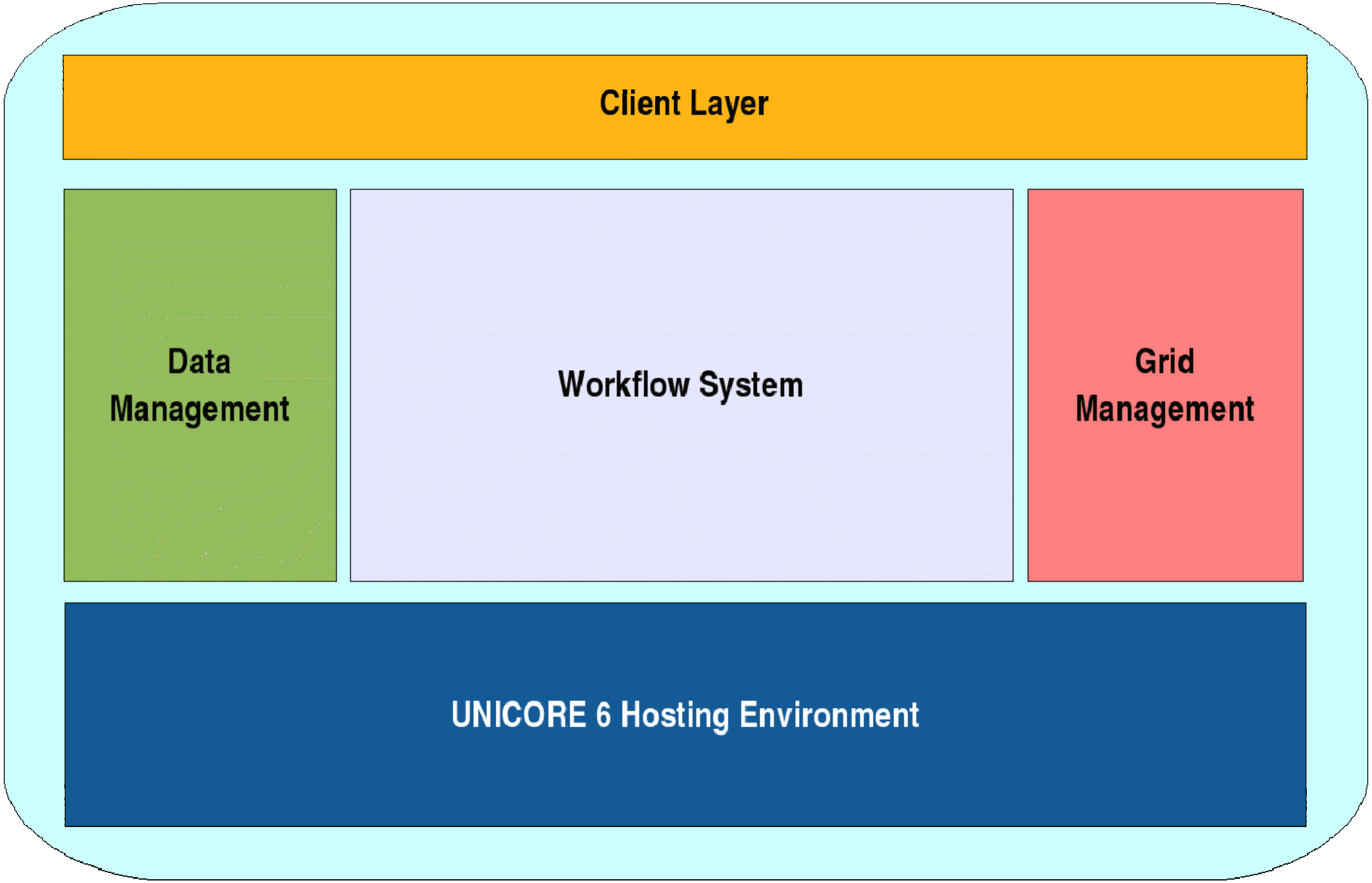
- Generic, flexible system for running workflow-centric, complex applications

  e.g. computational chemistry, supply-chain management
- Deals efficiently with data and knowledge
- Focused on end users
- Use cases: drug discovery, toxicity prediction, environmental risk assessment, QSAR, protein docking
- Based on UNICORE Grid middleware
- Web site: www.chemomentum.org

Chemomentum project overview

- 9 partners:
  - University of Warsaw, Poland (co-ordinator)
  - Research Centre Jülich, Germany
  - University of Tartu, Estonia
  - University of Technology Dresden, Germany
  - University of Ulster, United Kingdom
  - Istituto di Richerche Farmacologiche Mario Negri, Italy
  - University of Zurich, Switzerland
  - BioChemics Consulting SAS, France
  - TXT e-Solutions, Italy
- 30 month, started 01/07/2006

The big picture



Ambitions – Data Management

- Store data produced by workflows

  → need metadata to retrieve data later

  - General metadata, e.g. owner, dates, applications used, workflow description
  - Domain specific metadata, e.g. chemical structures inspected

- Calculation results should be reproducable

  → special attention to ensuring provenance of data

Ambitions – Data Management

- Handle files and meta information produced by Chemomentum
  - Store result files and meta /provenance information
  - Browse through stored data
  - Update and delete data
- Provide access to external data sources (e.g. chemical databases)
- Use ontologies to improve search results

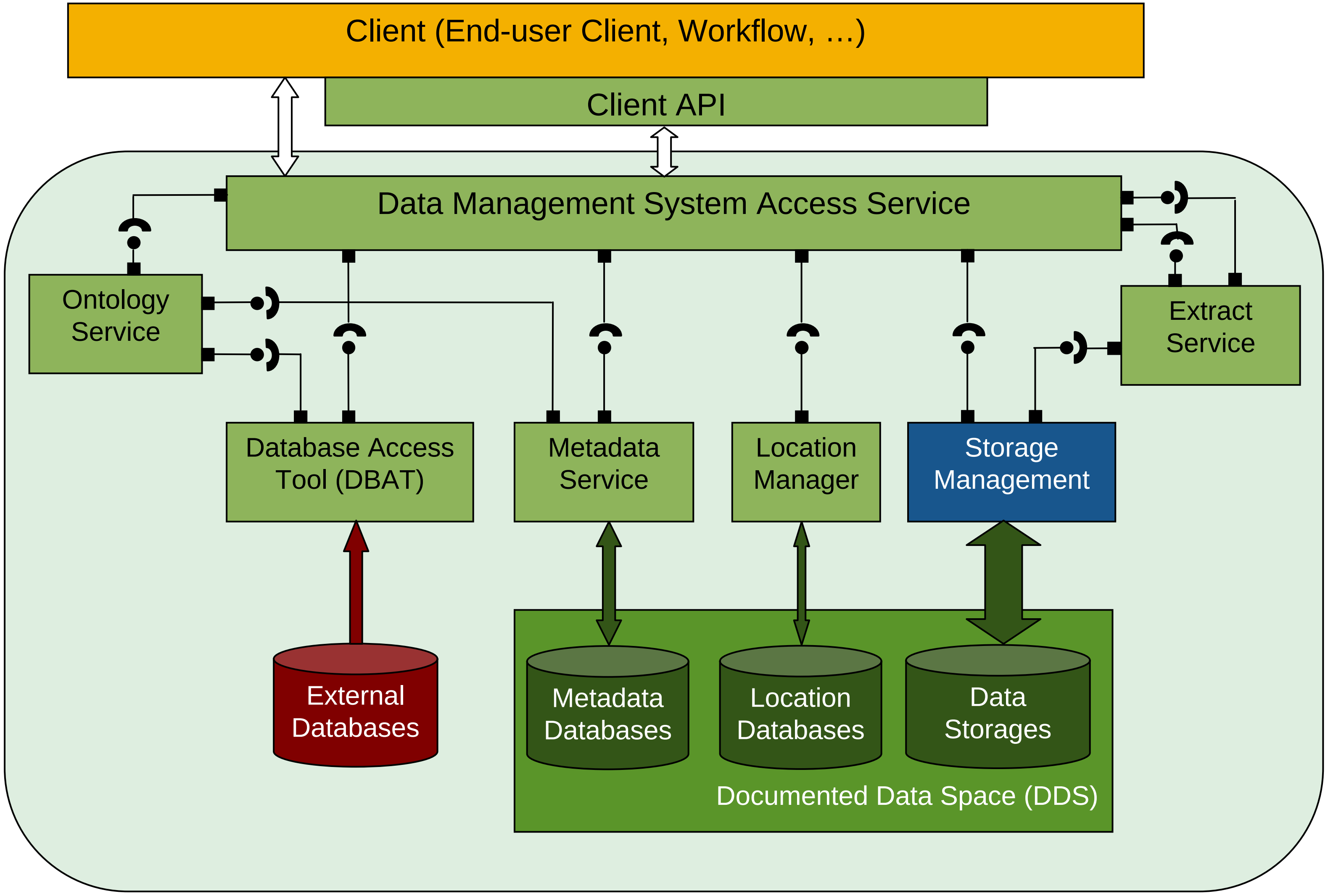
Features – Data Management

- Grid storage system
  - Data identified by globally unique logical name
    → global view of data
  - Data annotation with **extensible** meta/provenance data
  - Automatic metadata extraction
  - Distribution and replication
  - Seamless access to external data sources
  - Provide synonyms and unit conversion to improve request

Features – Data Management

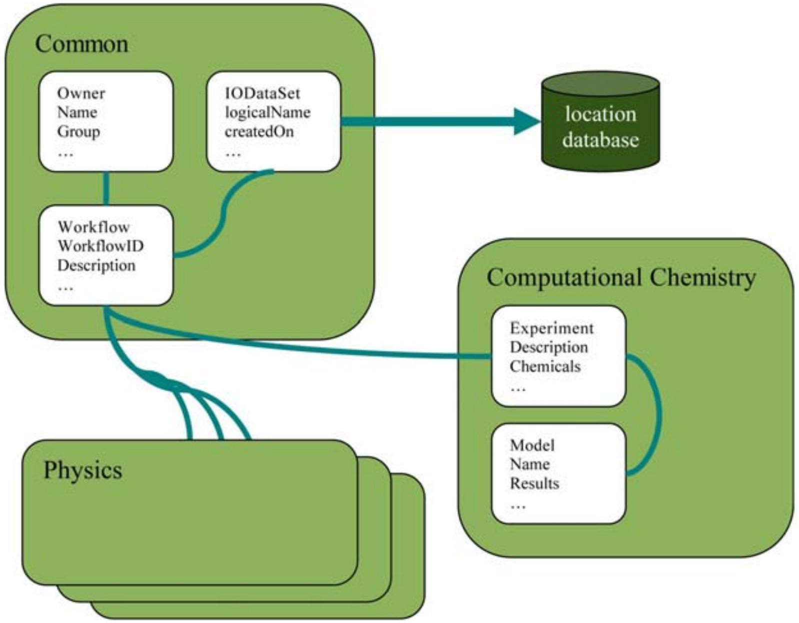
- Integrated into UNICORE/Chemomentum
  - Webservice based (using WSRFlite framework)
  - Workflow System uses data management to retrieve input files and store output files / meta information
- Integration into Chemomentum client
  - Query/browse through data and metadata
  - Manually upload/annotate/delete data and metadata
  - Administration

Components and Interfaces



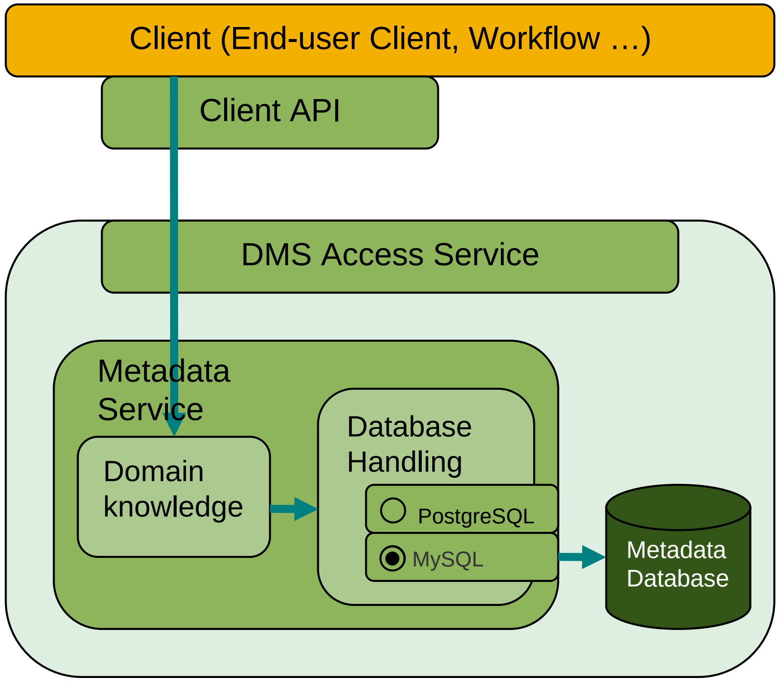
Metadata modelling

- Scientific administrator defines metadata schema for a scientific domain
- Contains tables and attributes
- Defines metadata properties:
  - Description
  - Data type
  - Unit
  - Provenance
  - Link to other attribute
  - …



Metadata modelling

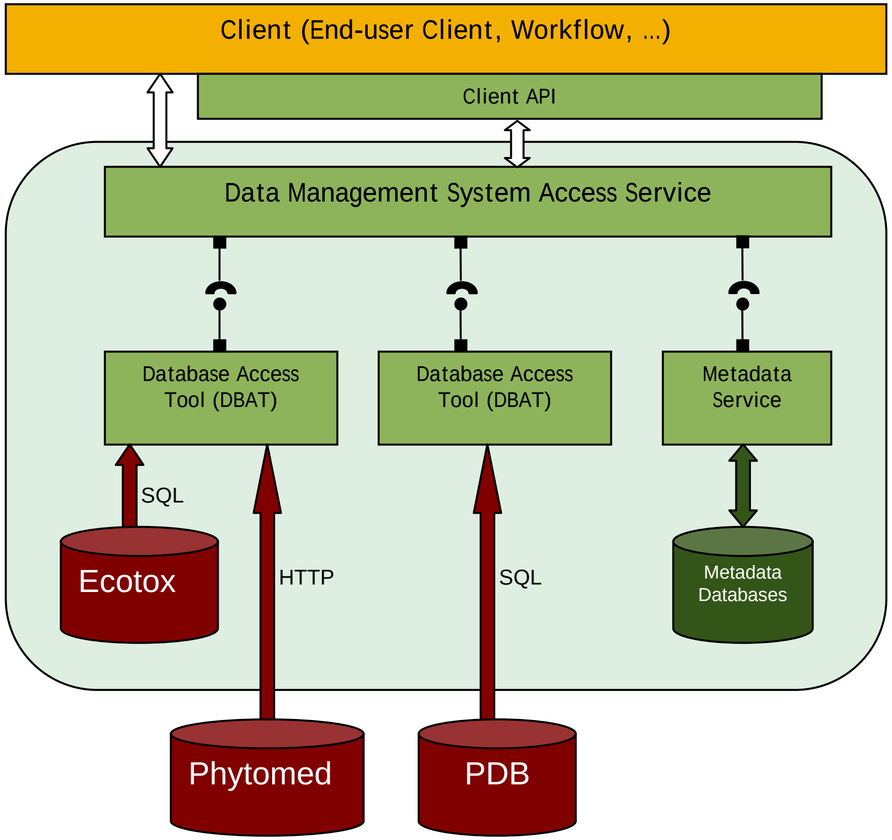
- Metadata exchanged in domain schema format
- Automatic query building using domain knowledge
- Pluggable database handlers for DMBS support
- GUI-based composition of new client views



Querying data and metadata

- Seamless access to external data sources: SQL databases, web services, Excel files, web forms

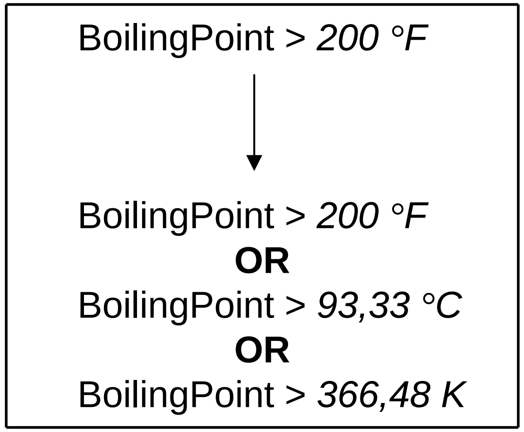
→ Access to data and metadata regardless of source, e.g. in workflow system



Client (End-user Client, Workflow, ...)
Client API
Data Management System Access Service
Database Access Tool (DBAT)
Database Access Tool (DBAT)
Metadata Service
SQL
HTTP
SQL
Ecotox
Metadata Databases
Phytomed
PDB

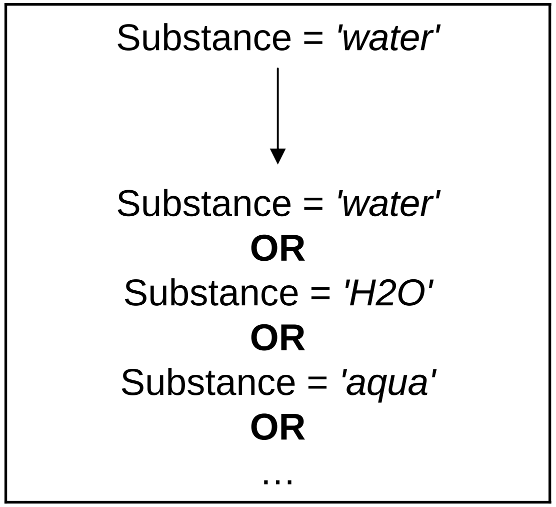
Querying data and metadata

- Automatic conversion of units in request and response
- Usage of external ontology services to broaden queries, e.g. synonyms from ChEbi



BoilingPoint > *200 °F*

BoilingPoint > *200 °F*
OR
BoilingPoint > *93,33 °C*
OR
BoilingPoint > *366,48 K*



Substance = *'water'*

Substance = *'water'*
OR
Substance = *'H2O'*
OR
Substance = *'aqua'*
OR
…

Substance	**BoilingPoint**
water	100 °C
arsenic	1137,2 °F
helium	-268.93 °C

Storing files and metadata

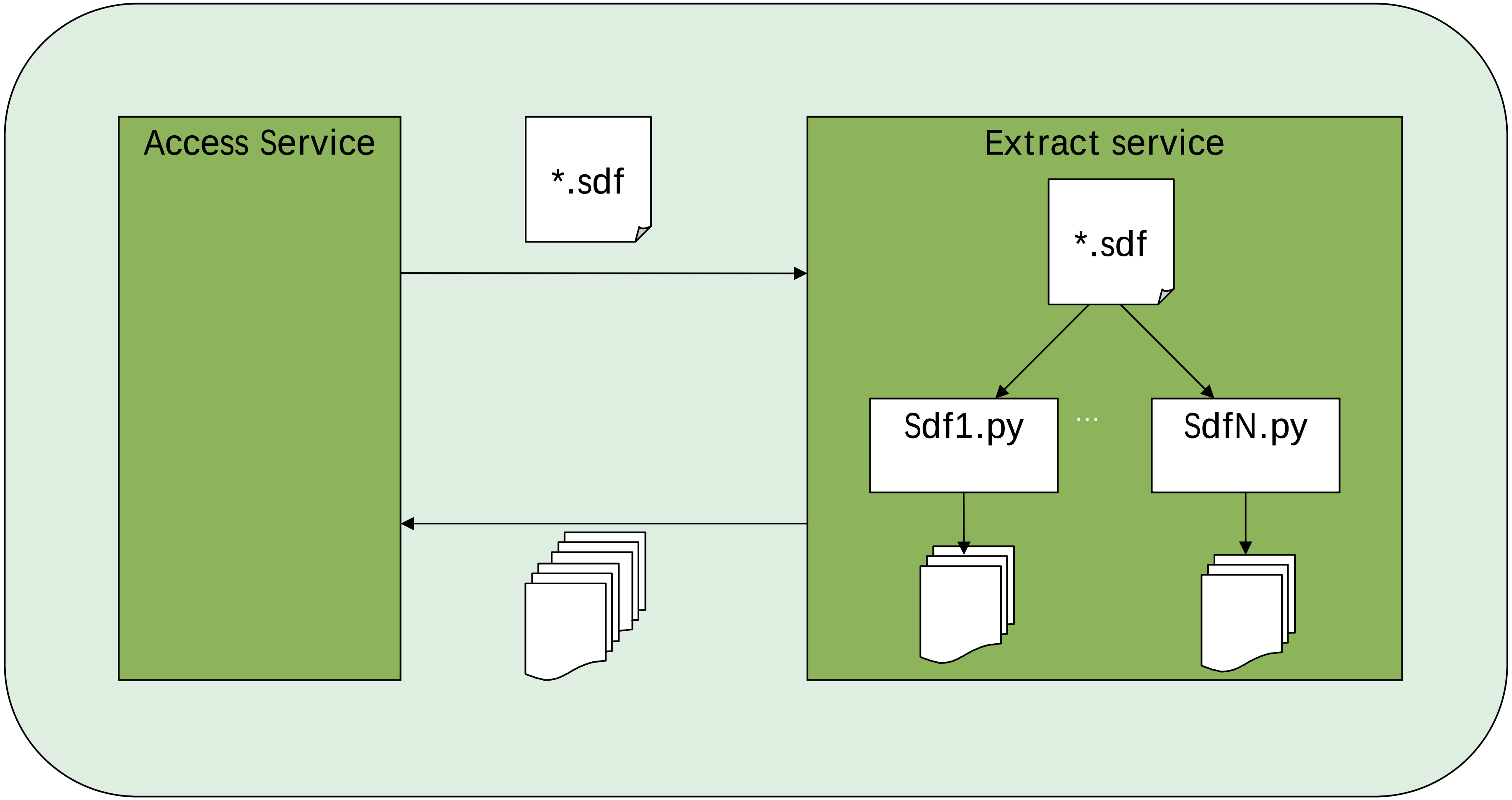
Example: Workflow system stores result of QSAR workflow

1. Store file on UNICORE6 Storage → URL to file
2. Register file with location manager → logical name
3. Execute necessary unit conversions on metadata
4. Store metadata include logical name
5. Extract metadata from file (e.g. Structure Data Format, SDF)
6. Store extracted metadata

Storing files and metadata

- Extract service:
  - Extraction logic in python scripts
  - Multiple extractors for single files possible
  - Uses metadata domain and file type to find matching extractors
  - Stores extracted metadata
  - e.g. create thumbnails from images, extract structure information from SDF file

Storing files and metadata

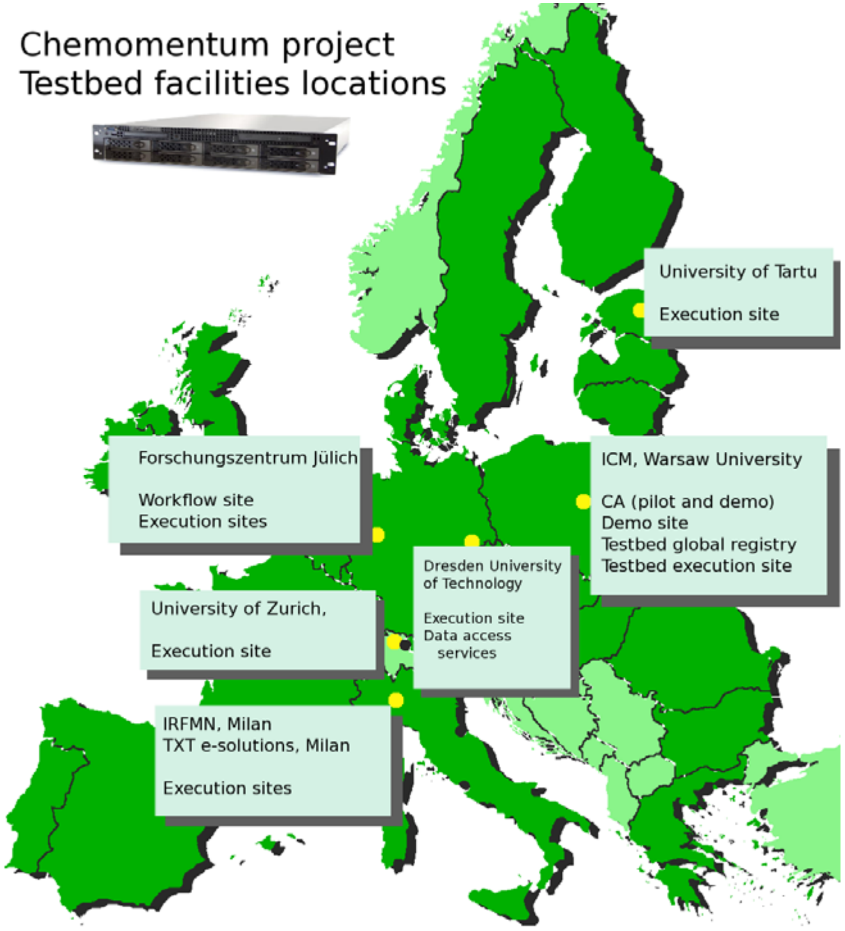


Security

- Uses UNICORE6 security infrastructure (X.509 certificates) to authenticate users
- XUUDB or Chemomentum VO management UVOS to authorise users
- Row-based access control lists for metadata and location information
- Metadata marked as provenance can only be modified/deleted by admin → provenance of calculation results

Testbed installation

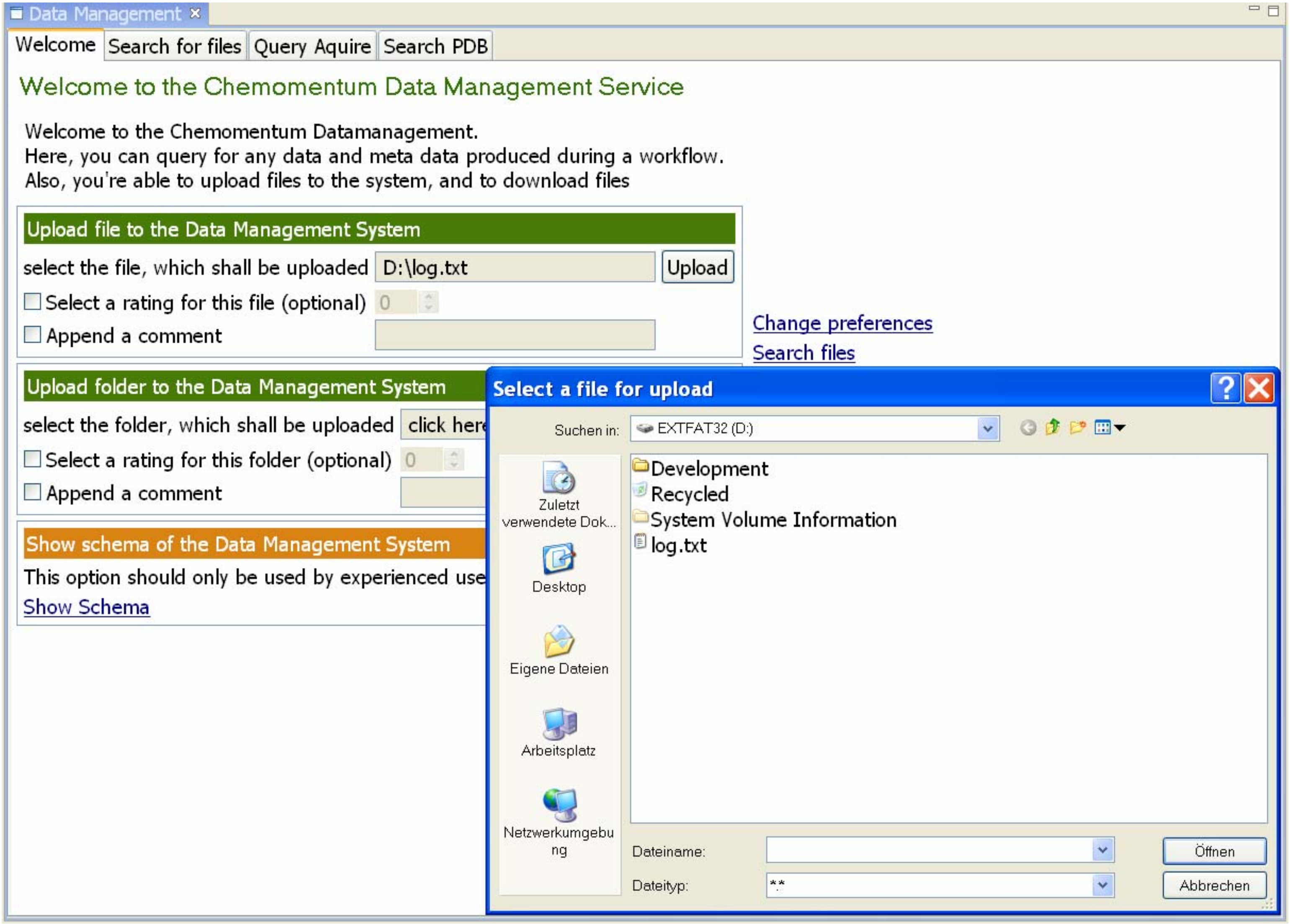
- Data Management System installed at TU Dresden
- Used by Workflow system to store workflow output and manage intermediate files



Client

- Based on Eclipse Rich Client Platform
- Query, store, update and delete data and metadata
- Administrative functions, e.g. edit/create domain schemas
- GUI-based composition of new client views using domain knowledge, e.g. generation of query forms
- Extension points to build own interaction possibilities (e.g. integration of other views for data visualisation)

Client: File upload



Welcome to the Chemomentum Data Management Service

Welcome to the Chemomentum Datamanagement.
Here, you can query for any data and meta data produced during a workflow.
Also, you're able to upload files to the system, and to download files

Upload file to the Data Management System

select the file, which shall be uploaded D:\log.txt Upload

□ Select a rating for this file (optional) 0

□ Append a comment

Upload folder to the Data Management System

select the folder, which shall be uploaded click here

□ Select a rating for this folder (optional) 0

□ Append a comment

Show schema of the Data Management System

This option should only be used by experienced use

Show Schema

Change preferences

Search files

Select a file for upload

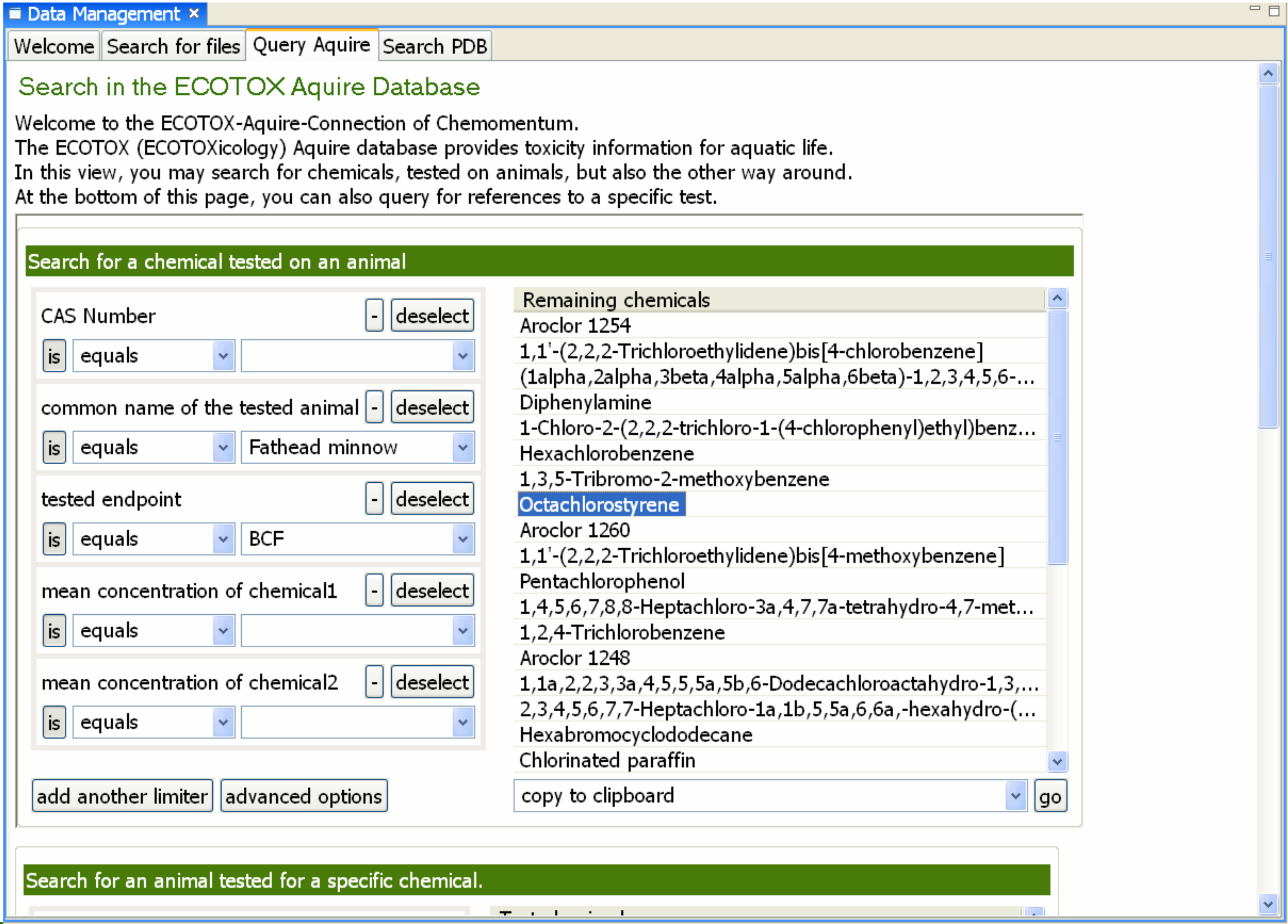
Suchen in: EXTFAT32 (D:)

Development
Recycled
System Volume Information
log.txt

Dateiname:

Dateityp: *.*

Client: Search aquire



Search in the ECOTOX Aquire Database

Welcome to the ECOTOX-Aquire-Connection of Chemomentum.
The ECOTOX (ECOTOXicology) Aquire database provides toxicity information for aquatic life.
In this view, you may search for chemicals, tested on animals, but also the other way around.
At the bottom of this page, you can also query for references to a specific test.

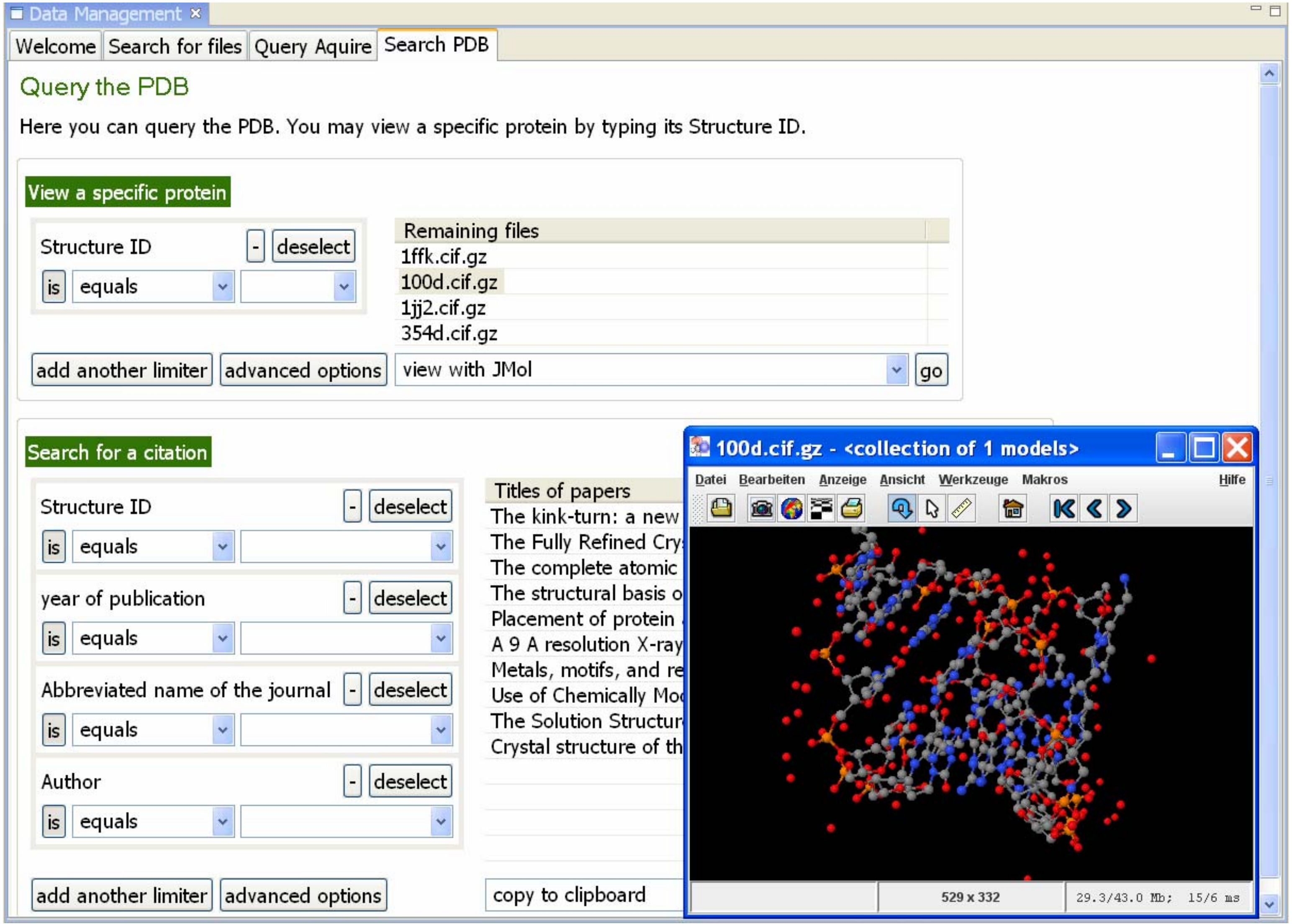
Search for a chemical tested on an animal

CAS Number: is equals
common name of the tested animal: is equals Fathead minnow
tested endpoint: is equals BCF
mean concentration of chemical1: is equals
mean concentration of chemical2: is equals

Remaining chemicals
Aroclor 1254
1,1'-(2,2,2-Trichloroethylidene)bis[4-chlorobenzene]
(1alpha,2alpha,3beta,4alpha,5alpha,6beta)-1,2,3,4,5,6-...
Diphenylamine
1-Chloro-2-(2,2,2-trichloro-1-(4-chlorophenyl)ethyl)benz...
Hexachlorobenzene
1,3,5-Tribromo-2-methoxybenzene
Octachlorostyrene
Aroclor 1260
1,1'-(2,2,2-Trichloroethylidene)bis[4-methoxybenzene]
Pentachlorophenol
1,4,5,6,7,8,8-Heptachloro-3a,4,7,7a-tetrahydro-4,7-met...
1,2,4-Trichlorobenzene
Aroclor 1248
1,1a,2,2,3,3a,4,5,5,5a,5b,6-Dodecachloroactahydro-1,3,...
2,3,4,5,6,7,7-Heptachloro-1a,1b,5,5a,6,6a,-hexahydro-(...
Hexabromocyclododecane
Chlorinated paraffin

Search for an animal tested for a specific chemical.

Client: PDB and JMOL



Query the PDB

Here you can query the PDB. You may view a specific protein by typing its Structure ID.

Remaining files
1ffk.cif.gz
100d.cif.gz
1jj2.cif.gz
354d.cif.gz

view with JMol

Titles of papers
The kink-turn: a new
The Fully Refined Cry
The complete atomic
The structural basis o
Placement of protein
A 9 A resolution X-ray
Metals, motifs, and re
Use of Chemically Mod
The Solution Structur
Crystal structure of th

100d.cif.gz - <collection of 1 models>

529 x 332

29.3/43.0 Mb; 15/6 ms

Thank you.