Managing and analyzing analytical chemistry data sets

Sana Tfaili

Sana Tfaili[1], Karima Rafes[2], Valérie Cantonny[3], Diem Bui Thi[1], Ali Tfayli[1] , Cécile Germain[4,5] and Pierre Chaminade[1]

[1]Lip(Sys)²- Chimie Analytique Pharmaceutique[§], Univ. Paris-Sud, Université Paris-Saclay, F-92290 Châtenay-Malabry, France.
[§](FKA EA4041 Groupe de Chimie Analytique de Paris-Sud)
[2]Bordercloud 245, quai de la Bataille de Stalingrad, F-92130 Issy-les-Moulineaux, France.
[3]IPS², CNRS, UMR 9213, Sciences du végétal : du gène à l'écosystème
[4]LRI- Laboratoire de Recherche en Informatique, Bât 650, Rue Noetzlin, F-91190 Gif-sur-Yvette, France.
[5]LAL- Laboratoire de l'Accélérateur Linéaire, Bât 200, voie de la faculté, F-91440 Bures-sur-Yvette, France.

1. **Lip(Sys)² data sets.**

2. **Participation to CDS 1.0: objective, accomplishments and obstacles.**

3. **What are we seeking for in CDS 2.0?**

4. **Importance and issues for project continuity.**

1. Lip(Sys)² data sets

Different Analytical measurement techniques	Data processing	Domain scientists
Separation techniques, mainly chromatographic	Data (pre)-processing	Data providers
Coupled mass spectrometry techniques (LC, GC, GCxGC/MS)	Multivariate analysis	Lipid analysis
Vibrational spectroscopies (infrared, near infrared and Raman)	Chemometric techniques	

2. Participation to CDS 1.0: objective, accomplishments and obstacles.

Database creation and management

Paris-Saclay Center for Data Science

Improving data analysis: computing power (Cloud)

ERM CLOUD@VIRTUALDATA

Management of Data for the Analysis of Lipids, Metabolites and Isotopes 'MoDALMI'

- **Data storage**
- **Ontology, metadata, common format for data fusion (ex:.aia, .cdf)**
- **Automate data conversion and pre-treatment, open access**

✓ **1 year engineer** (plugins for data conversion)

✓ **Support from BorderCloud** (Linked data technologies)

✓ **Building DAAP Platform**

✓ **Working on a user friendly interface** (VRAC)

✓ **40 Tb storage** (ERM, AAP attractivité)

✓ **Ramp on drug cancer for drug identification and quantification**

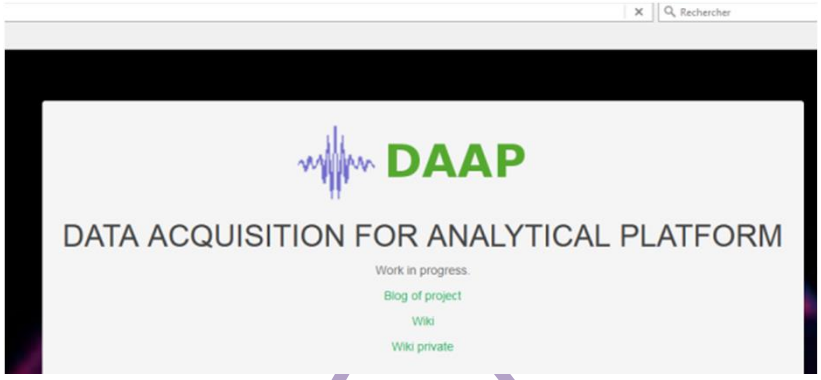
DAAP: DATA ACQUISTION FOR ANALYTICAL PLATFORM

DAAP

✓**Automating scientific workflows and building an open database platform for chemical analysis metadata**

Defining DAAP domain

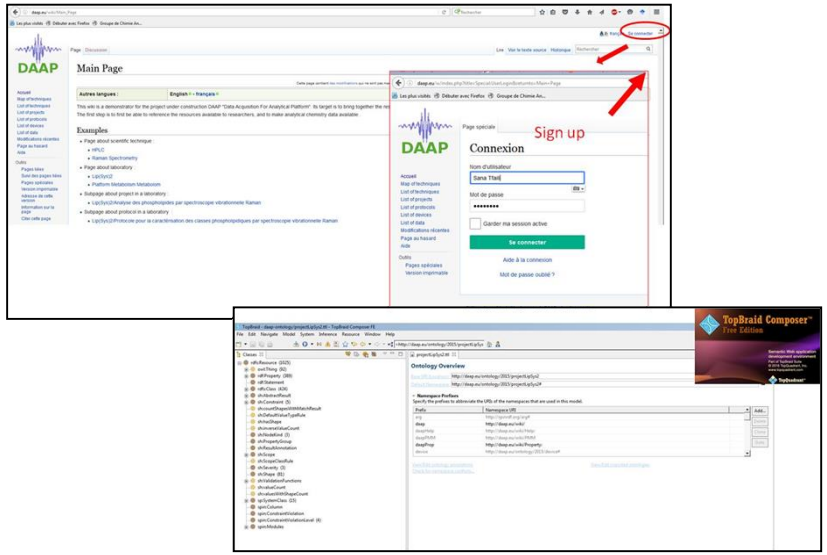
http://daap.eu/

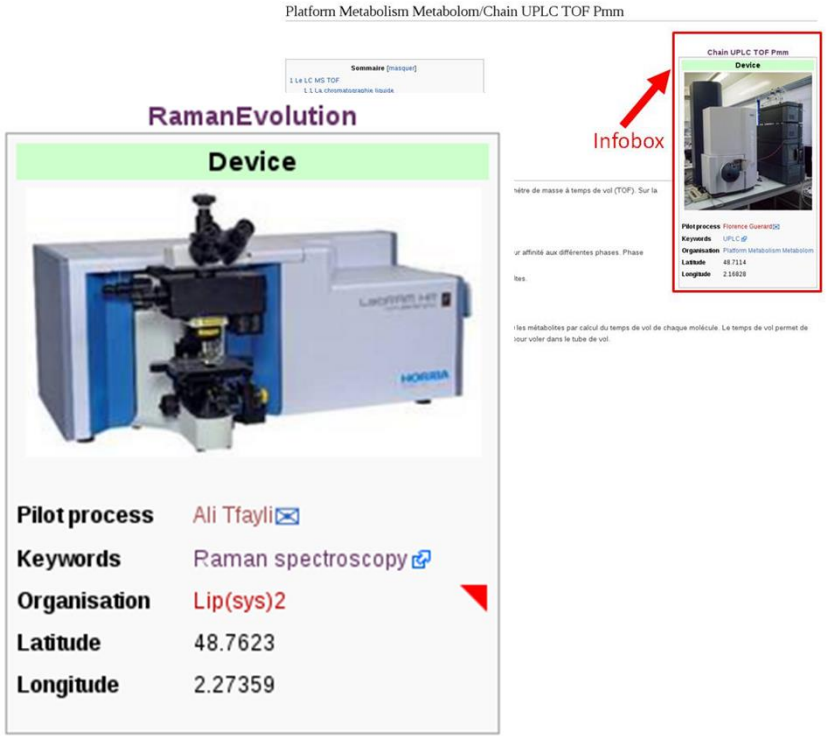


Public Wiki **Private Wiki**

Workflow steps

Step 1: Ontology defining and Data description in WikiDAAP pages.





Step 2: Data hosting.

ONTOLOGY Hosting
via SPARQL access point within Paris Sud University

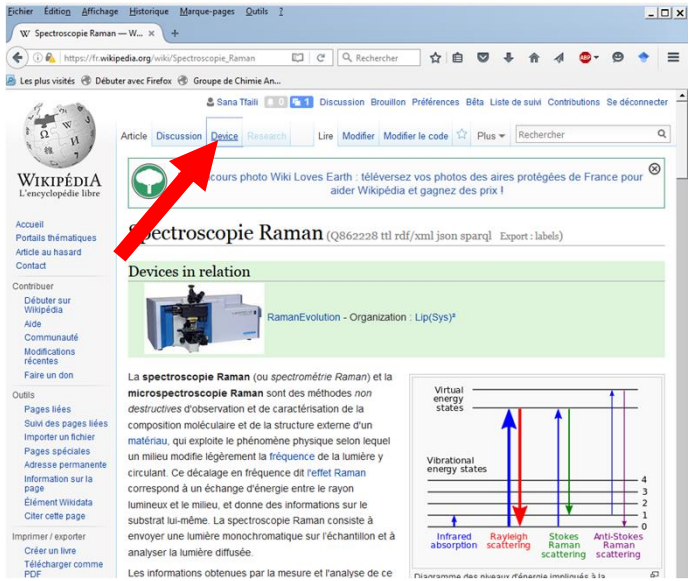
IRI were used to describe the IDs of all concepts.
The IRI links on wiki public pages.
RDF files that describe the ontology hosted in:
http://daap.eu/ontology/2015

The infobox in the wiki compares information between the formal ontology defined with TopBraid Composer

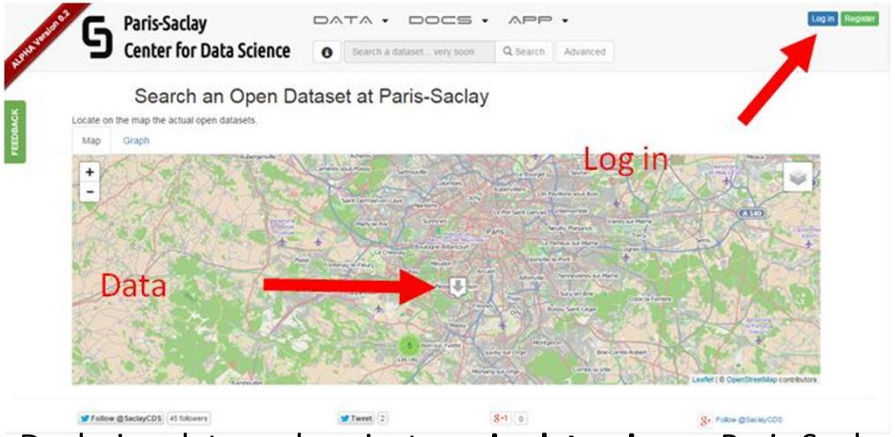
divergence (red flag)

Workflow steps

Step 3: linking the ontology to linked data technologies.

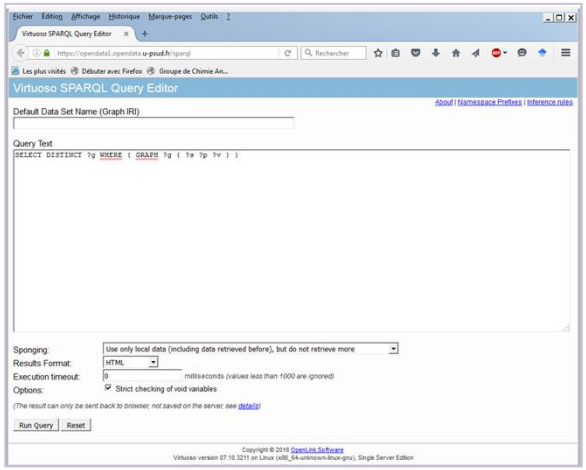


Definition the ontology of available instrument of measurements. Wikipedia keywords permitted to link each measurement technique by using the same keyword in the infobox of wikiDAAP.

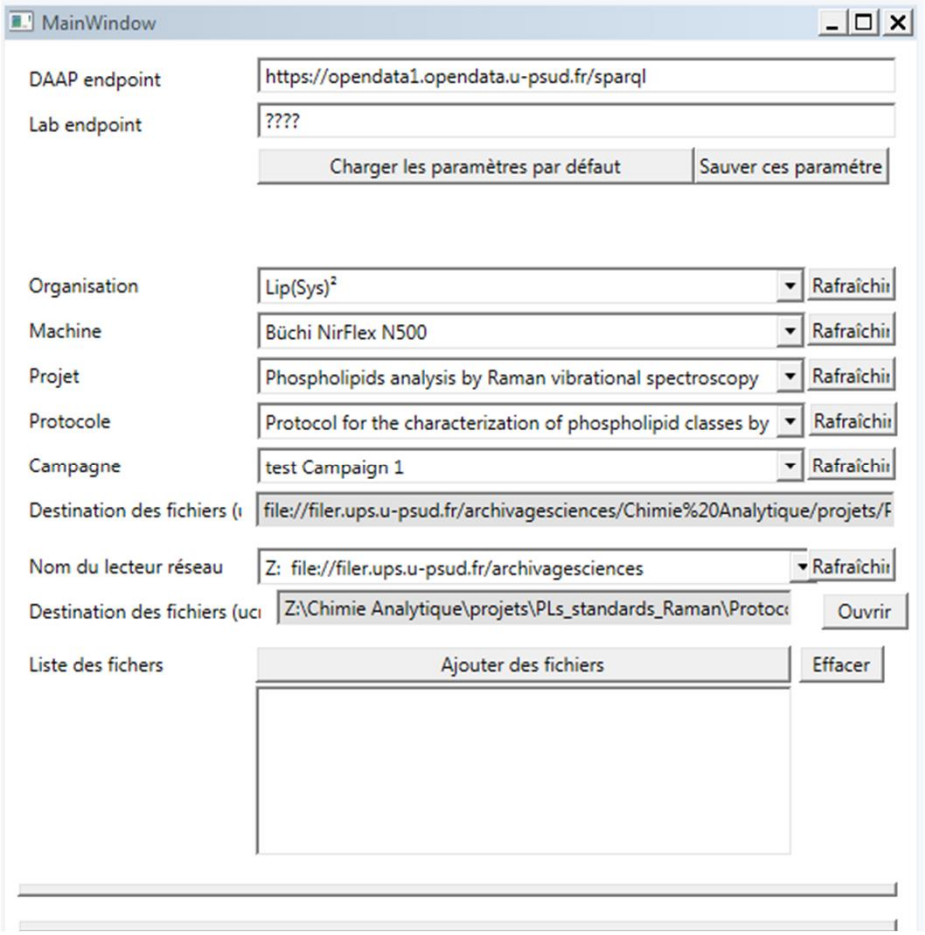


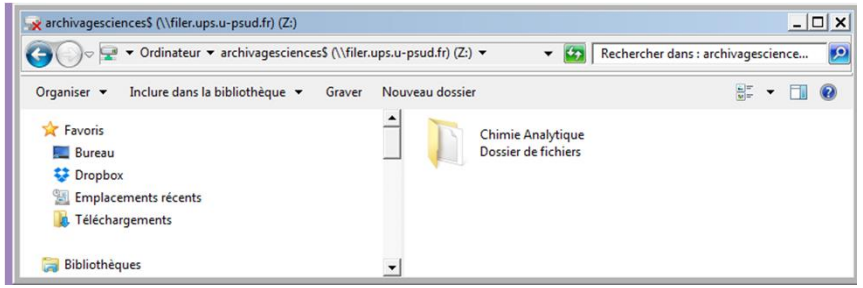
Declaring data and projects on **io.datascience** Paris Saclay.
https://io.datascience-paris-saclay.fr

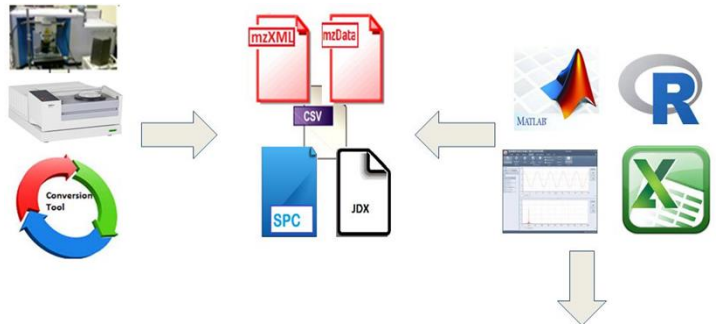
Perform queries

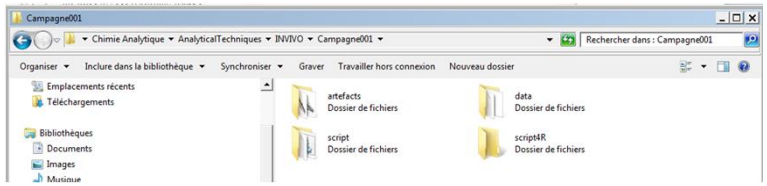


Virtual Research Environment for Analytical Chemistry
VRAC



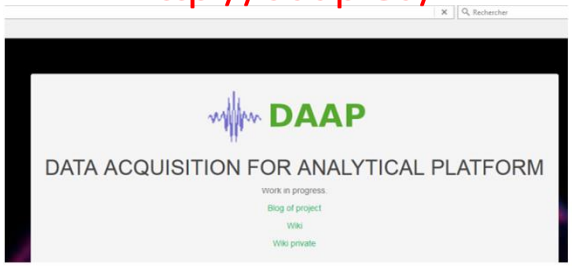


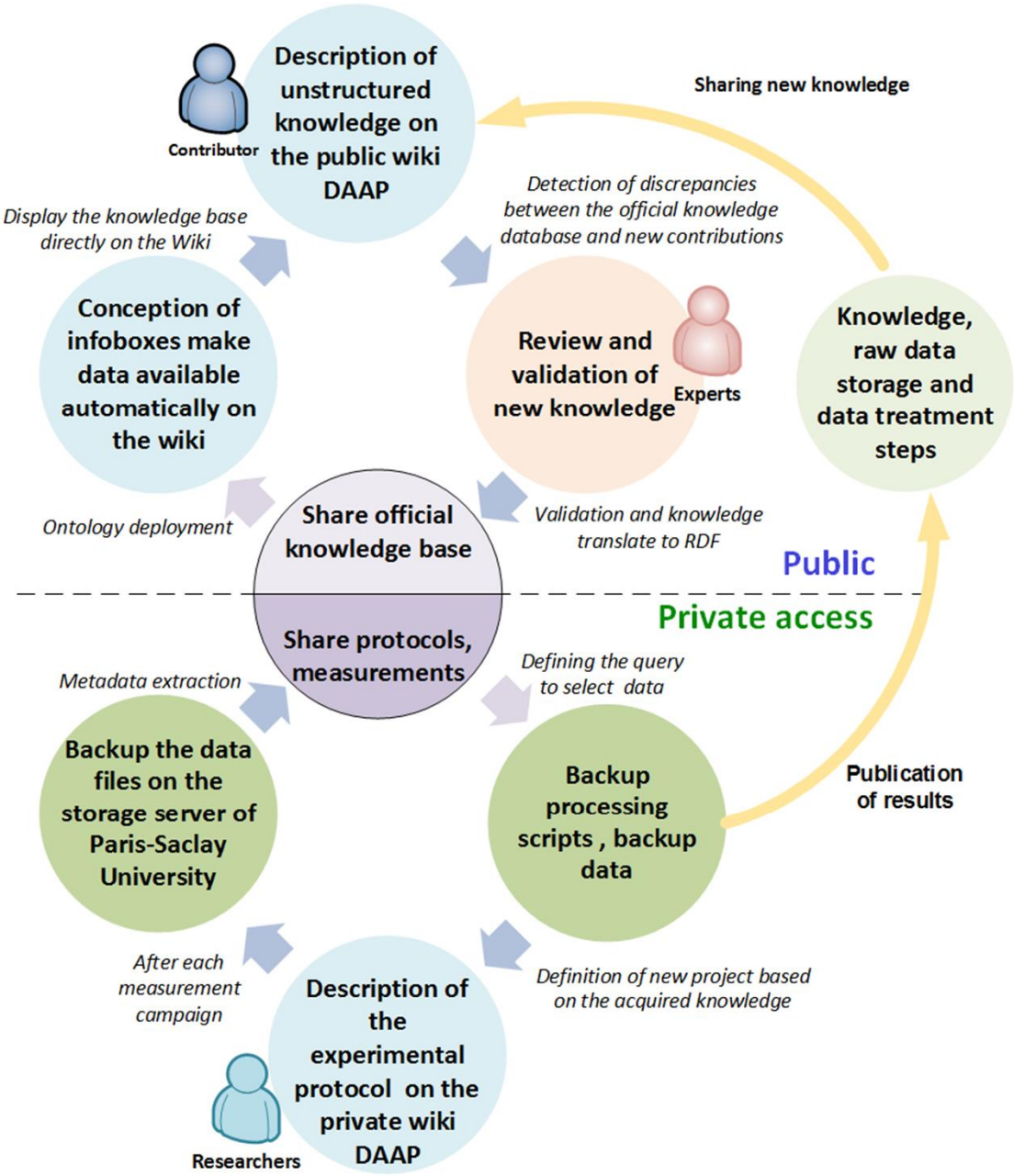


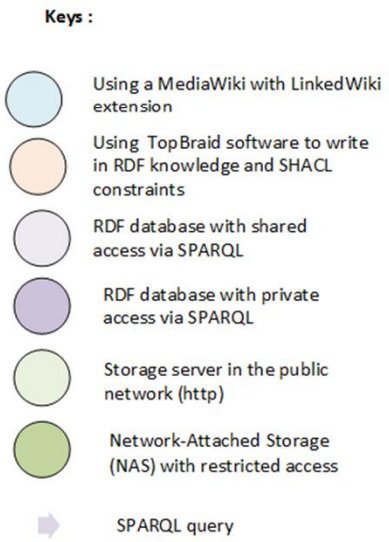


The Workflow

http://daap.eu/







CDS 2.0 project

3. What are we seeking for in CDS 2.0?

- *Enrich our database, define the ontology for the projects and data sets*
- *Optimize some plugin for data conversion*
- *user friendly software VRAC*
- *Analyze conjointly data sets of various origins*
- *new kinds of computational analyses*

Data scientists support.

Training (ontology, SPARQL queries and Wikimedia).

CDS 2.0 project

3. What are we seeking for in CDS 2.0?

- *Qualitative identification and quantitative prediction (a RAMP was organized around cancer drugs, Laetitia Le)*
- *Correction of the response of two universal detectors used in chromatography techniques (peak estimation and modeling)*
- *Data fusion between mass spectra obtained in positive and negative ionization mode*
- *Analysis of spectral images obtained by vibrational spectroscopy (Raman and Infrared spectroscopy)*

CDS 2.0 project
4. Importance and issues for project continuity

- *Data providers, Open access data sets (correctly described)*
- *Workflow is transposable to other research structures and could be interesting for industries in pharmaceutical domain.*
- *get external funding by combining in the same project between the defined workflow and health-related research topics*

Thank you!