PeakForest: a spectral database and its toolbox, dedicated to the Metabolomics' community

Nils Paulhe, Christophe Duperier, Marie Lefebvre, Pierrick Roger-Mele, Christophe Junot, Etienne Thévenot, Franck Giacomoni, Estelle Pujos-Guillot

PEAKFOREST: A SPECTRAL DATABASE AND ITS TOOLBOX, DEDICATED TO THE METABOLOMICS' COMMUNITY

NILS PAULHE[1], CHRISTOPHE DUPERIER[1], MARIE LEFEBVRE[4], DANIEL JACOB[4], PIERRICK ROGER-MELE[2], CHRISTOPHE JUNOT[3]†, ÉTIENNE THÉVENOT[2]†, FRANCK GIACOMONI[1] AND ESTELLE PUJOS-GUILLOT[1]†

1: PFEM, UMR1019, INRA, CENTRE CLERMONT-FERRAND-THEIX, 63122 SAINT GENÈS CHAMPANELLE, FRANCE
2: CEA, LIST, LABORATORY FOR DATA ANALYSIS AND SYSTEMS' INTELLIGENCE (LADIS), METABOHUB PARIS, F-91191 GIF-SUR-YVETTE, FRANCE
3: LABORATOIRE D'ETUDE DU MÉTABOLISME DES MÉDICAMENTS, DSV/IBITEC-S/SPI, METABOHUB PARIS, CEA SACLAY, 91191 GIF-SUR-YVETTE, FRANCE
4: BORDEAUX METABOLOME FACILITY, BORDEAUX FUNCTIONAL GENOMICS CENTER, METABOHUB PMB, INRA, 33140 VILLENAVE D'ORNON, FRANCE.

†: METABOHUB WOKPACKAGE LEADER

interested?

Contact us!

Introduction

The French National Infrastructure in Metabolomics & Fluxomics (**MetaboHUB**) provides a wide range of services for the **untargeted metabolomic analysis of biofluids, tissue samples, cell extracts and environmental matrices**, through its network of platforms. To increase the **high-throughput annotation and *de novo* identification** of metabolites, the consortium is producing **high quality reference MS, MS/MS and NMR annotated spectra**.

The **PeakForest** spectral library is being developed to allow **centralizing** and **sharing spectral data** from **reference compounds** and **metabolic profiles**, as well as **annotation tools**. The originality of the PeakForest project is based on the database structure and on spectral data quality. A **system managing the data life cycle** (import, enrich then re-use data in an annotation workflow) built on the multidisciplinary expertise of biologists, chemists, analysts and software designers, is proposed.

PeakForest

Data and services: PeakForest v1.1 already provides basic functionalities like a database of metabolites and their reference and profile spectra but also **querying tools and data visualization** plugged on it.

It provides templates and **user-friendly interfaces** in order to **collect metabolites and spectral data** (peak lists and **metadata**) from proton 1D-NMR and LC-MS instruments.

Spectra **annotations** are collected *via* templates of reference spectra. In mass spectra we gather all peaks (parent ion, isotopes, adducts and in-source fragments); In NMR spectra, each peak is labeled to a molecule atom number (figure 1). For experimental data we added the possibility to refer parts of a fingerprint matching an unknown compound to a "putative compound".

Originalities: Every data-provider can improve the database quality by adding "curation messages" to compounds or spectra sheets. Then a **curator** user is informed of the potential error and has extra rights to correct or enrich existing data.
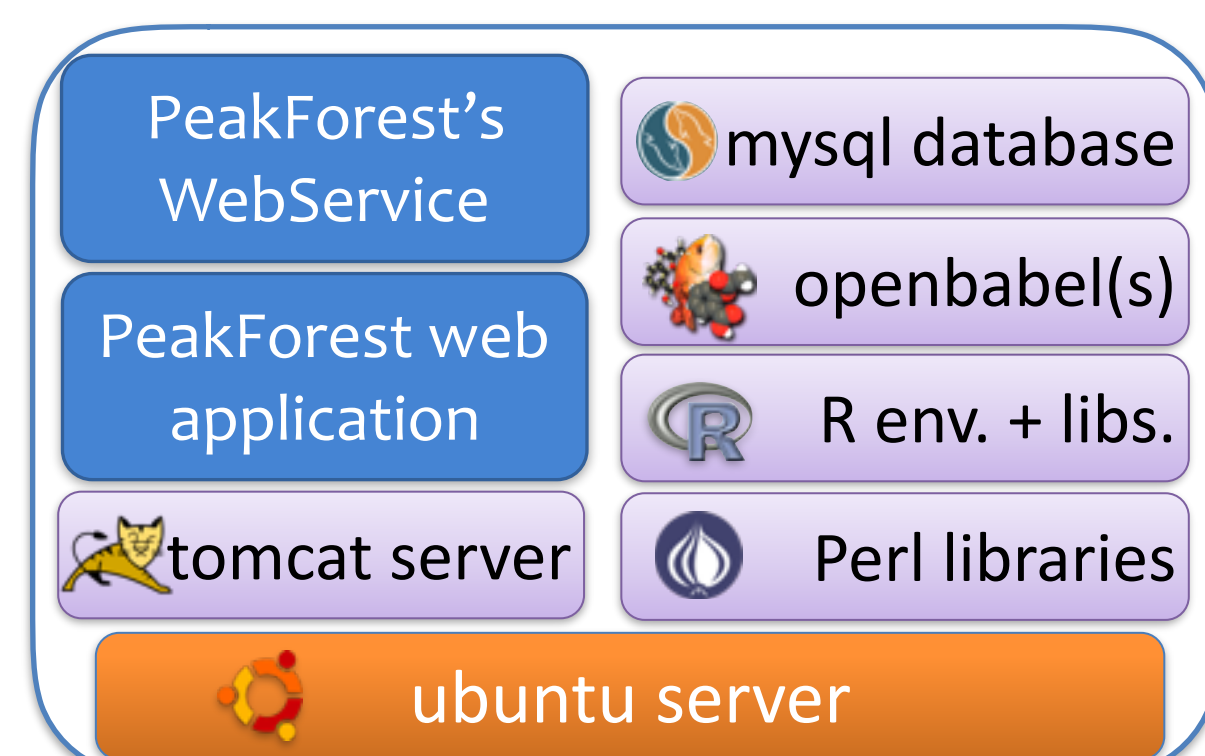
Computational specifications: PeakForest core API is split in different java libraries used in two major projects: the PeakForest Spectral Database, a user-friendly **web-application** (reference instance at peakforest.org) and **REST webservices**, a bot-friendly API (demo and documentation at doc-ws.peakforest.org).

The host server requires a custom configuration in order to support **third part tools plugged** on it like "peak-matching" tools or the chemical format converter library "OpenBabel" (OLBoyle *et al.* 2011[a]); *cf* figure 2.
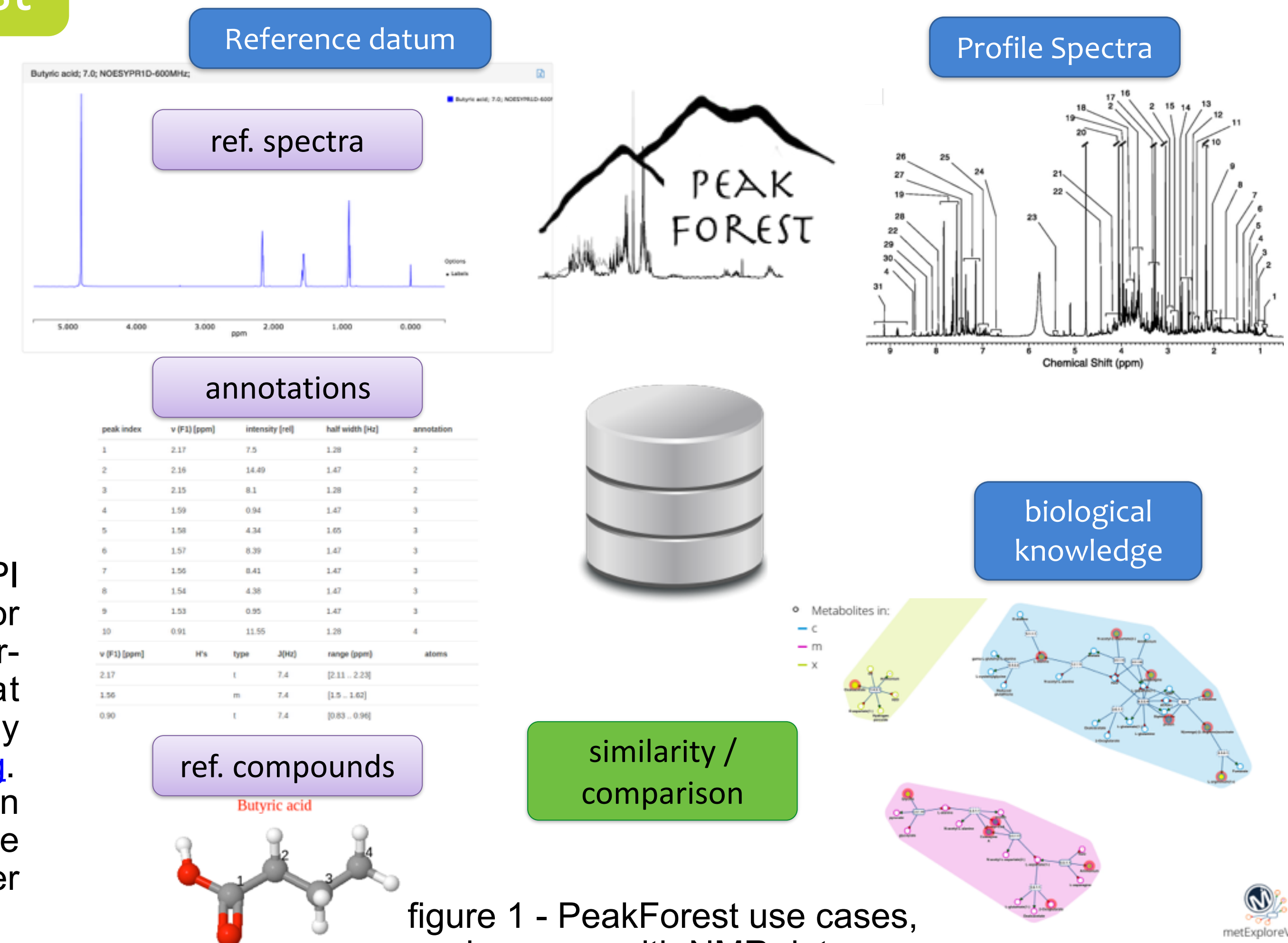
PeakForest's WebService | **mysql database**
PeakForest web application | **openbabel(s)**
tomcat server | **R env. + libs.**
| **Perl libraries**
ubuntu server

figure 2 - host server configuration

Reference datum

ref. spectra

annotations

ref. compounds
Butyric acid

Profile Spectra

biological knowledge

similarity / comparison

figure 1 - PeakForest use cases, showcase with NMR datum

toolbox

Hosted tools and open toolbox: PeakForest's online querying and peakmatching tools (lcmsmatching by Pierrick Roger-Mele and NMR peakmatching by Marie Lefebvre) are available to all users. Homemade tools can also be plugged into PeakForest thanks to PeakForest REST API: it provides **different querying services** in order to extract spectra subsets ready to **be used by any MS / chemical shift matching algorithms** (*cf* figure 3).

The main purpose of this bot-friendly interface is to open stored datum and use them as a resource for **computer aid annotation tools** in **high-throughput metabolomics** context.

A proof of concept has been developed through a **Galaxy** tool, soon available *via* workflow4metabolomics.org **W4M platform** (Giacomoni *et al.* 2015[b]), the collaborative portal dedicated to metabolomics data processing analysis and annotation for the Metabolomics community. We are working on the possibility to call any instance of PeakForest, even private, through this galaxy tool, *via* a token management system.
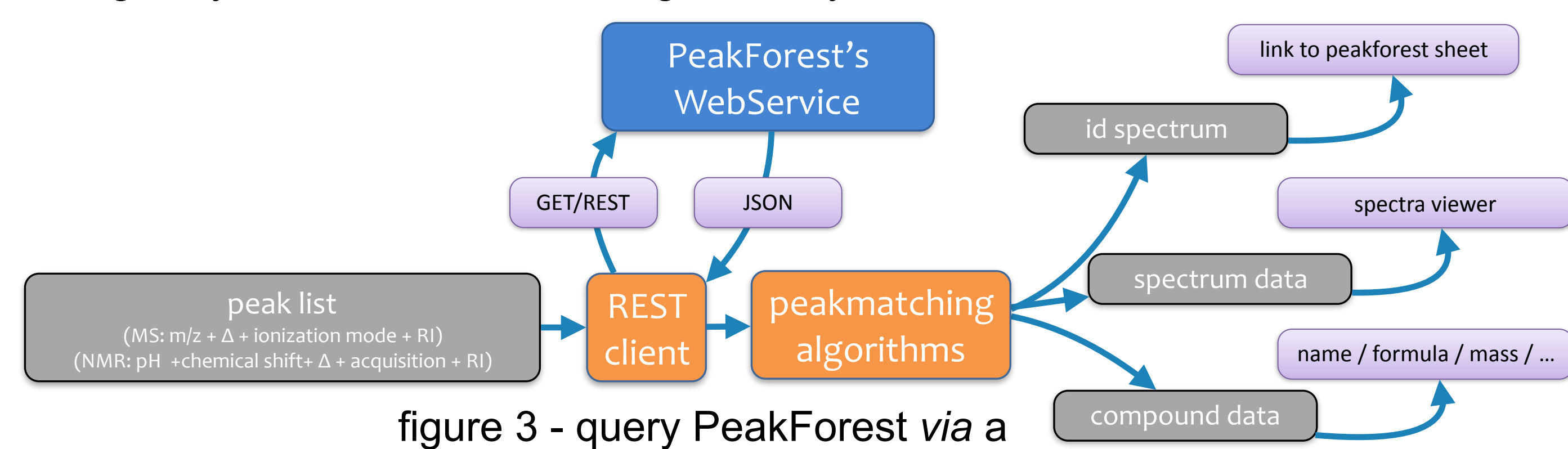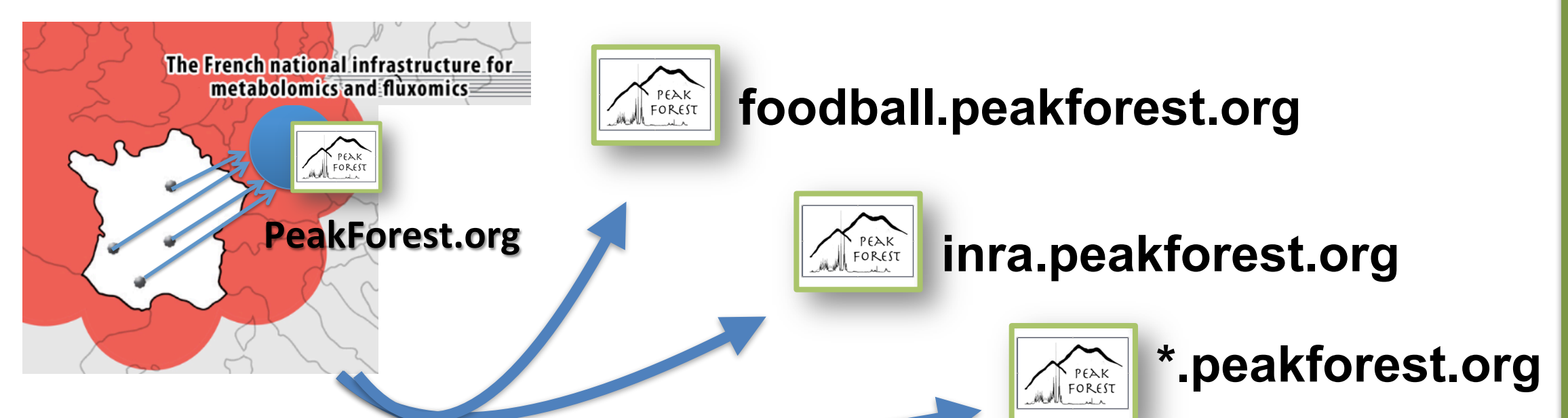
PeakForest's WebService

peak list
(MS: m/z + Δ + ionization mode + RI)
(NMR: pH + chemical shift + Δ + acquisition + RI)

GET/REST | JSON

REST client | peakmatching algorithms

link to peakforest sheet
id spectrum
spectra viewer
spectrum data
name / formula / mass / ...
compound data

figure 3 - query PeakForest *via* a REST client

Cross service exchanges between databases: **webservices** are used to link MetaboHUB's tools for mutual enrichment! *e.g.* we can provide chemical informations on **MetExplore**'s networks (*cf* poster 40 " MetExplore: handling genome scale metabolic networks online" by Jourdan *et al.*) from **PeakForest** compounds sheets and use **MetExplore's** mapping coverage statistics in **Peakforest** (Merlet *et al.*, 2016[c]).

what about YOUR own PeakForest?

PeakForest.org is MetaboHUB's showcase to provide its reference datum to the whole metabolomics community. But PeakForest is also a project with a standardized format available through its Java environment where you can **create your own database**! Thanks to virtualization technologies we are able to quickly deploy and host new instances of this tool.

Furthermore the PeakForest development team is available to **maintain and update** new releases of PeakForest on your own instance! If you are interested, please **contact us** :-) ! <contact@peakforest.org>
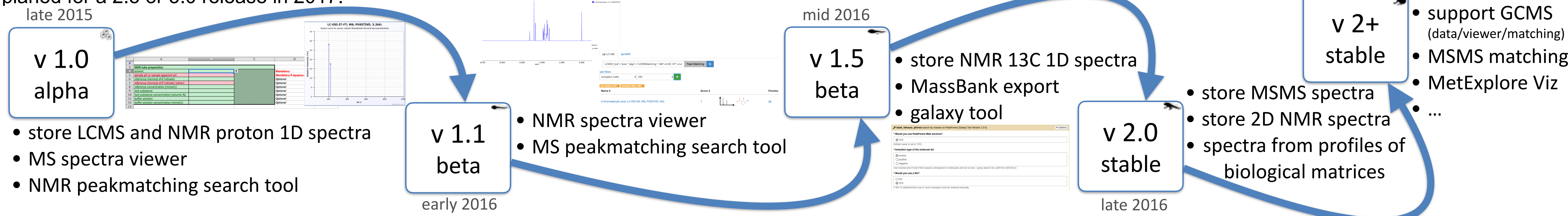
The French national infrastructure for metabolomics and fluxomics

PeakForest.org

foodball.peakforest.org

inra.peakforest.org

*.peakforest.org

Why? it's relevant to create private sub instances of PeakForest **specialized** in a scientific domain (*e.g.*: the "foodball" project focused on food metabolome) or to store and process **unpublished results**.

Data agreement in return: the main instance of PeakForest may be enriched thanks to these sub-datasets after result publication or in order to **cover wide metabolomics fields**.

Perspectives

PeakForest version 2.0: new functionalities including, **2D-NMR** and **LC-MS/MS** spectra management are scheduled. The PeakForest toolbox is ready to host services compliant with these data. At the **end of 2016**, this version of PeakForest will provide **storage and services for annotation of biological reference matrices**. Long-term functionalities like **GC-MS spectra management** and **MetExploreViz** integration (cf poster 47 "MetExploreViz: Visualisation of metabolic online" by Chazalviel *et al.*) are planed for a 2.5 or 3.0 release in 2017.

late 2015
v 1.0 alpha
• store LCMS and NMR proton 1D spectra
• MS spectra viewer
• NMR peakmatching search tool

v 1.1 beta
• NMR spectra viewer
• MS peakmatching search tool
early 2016

mid 2016
v 1.5 beta
• store NMR 13C 1D spectra
• MassBank export
• galaxy tool

v 2.0 stable
• store MSMS spectra
• store 2D NMR spectra
• spectra from profiles of biological matrices
late 2016

v 2+ stable
• support GCMS (data/viewer/matching)
• MSMS matching
• MetExplore Viz
• ...

Special thanks to all **MetaboHUB** members who have acquired and deposited NMR or MS spectra into PeakForest.

references:
[a]: OLBoyle, Noel M., et al. "Open Babel: An open chemical toolbox." *J Cheminf* 3 (2011): 33.
[b]: Giacomoni, Franck, et al. "Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics." *Bioinformatics* 31.9 (2015): 1493-1495.
[c]: Merlet, Benjamin, et al. "A Computational Solution to Automatically Map Metabolite Libraries in the Context of Genome Scale Metabolic Networks." *Frontiers in molecular biosciences* 3 (2016).

http://www.clermont.inra.fr/plateforme_exploration_metabolisme