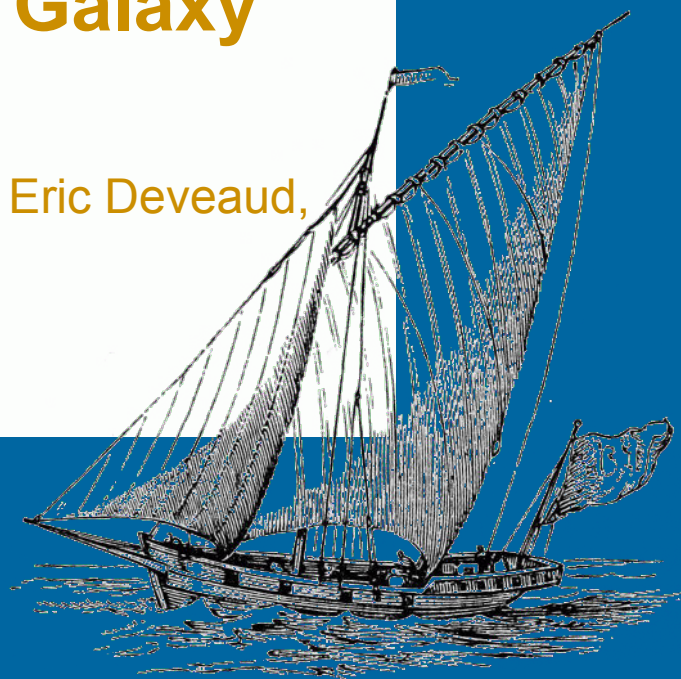


ReGaTE, Registration of Galaxy Tools in Elixir

Olivia Doppelt-Azeroual, Fabien Mareuil, Eric Deveaud,
Matus Kalas, and Hervé Menager

07/07/2015



Plan

INTRODUCTION

Original questions

PART 1

Elixir Registry

PART 2

ReGaTE

CONCLUSION

& Perspectives



INTRODUCTION

Original questions



1.1 Original Questions

(A regate is a boat race in French)



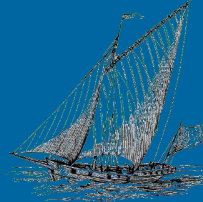
Data analyst

metagenomic RNAseq
On what kind of project does he work ?
Agent detection proteomics

BAM fastQ VCF
What kind of data does he have/want ?
fasta Krona report

A Galaxy instance A cluster with admins
What ressources does he have access to ?
Not much A virtual machine

1.2 Start answering, with a galaxyst ?



if using a Galaxy with admin knowledge → Search in a Toolshed and build analysis workflows but ...

WARNING !!

1. HOW ARE TOOLS CLASSIFIED ?
2. NOT ALL TOOLS ARE IN GALAXY

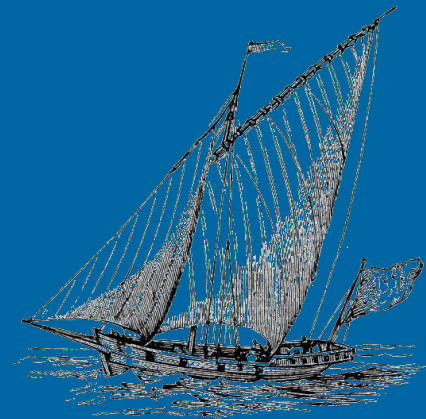
The screenshot shows the Galaxy Tool Shed interface. The top navigation bar includes links for Applications, Bookmarks, and various project folders. The main content area is titled 'Galaxy Tool Shed' and displays a table of repositories categorized by type. The table has three columns: Name, Description, and Repositories. The categories listed on the left include Assembly, Convert Formats, Data Management, Data Source, Fastq Manipulation, Genomic Interval Operations, Graphics, Metagenomics, Nebula, Next Gen Mappers, Ontology Manipulation, phylogeny, Picard tools, Proteomics, SAM, Sequence Analysis, SNP Analysis, Statistics, Text Manipulation, URG, VCF, Visualization, and xml test.

Name	Description	Repositories
Assembly	Tools for working with assemblies	2
Convert Formats	Tools for converting data formats	15
Data Management	Tools for managing data	6
Data Source	Tools for retrieving data from external data sources	1
Fastq Manipulation	Tools for manipulating fastq data	7
Genomic Interval Operations	Tools for operating on genomic intervals	2
Graphics	Tools producing images	1
Metagenomics	Tools enabling the study of metagenomes	2
Nebula	Nebula is a web service provided by Institut Curie and powered by Galaxy which allows users (Bioinformaticians as far as Biologists) to analyze their ChIP-seq data.	13
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	28
Ontology Manipulation	Tools for manipulating ontologies	
phylogeny	tools for phylogeny	1
Picard tools	Galaxy wrappers for the Picard SAM/BAM manipulation tools. (Pasteur version)	2
Proteomics	Tools enabling the study of proteins	
SAM	Tools for manipulating alignments in the SAM format	18
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	15
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	2
Statistics	Tools for generating statistics	4
Text Manipulation	Tools for manipulating data	3
URG	URG tools	1
VCF	Tools for manipulating vcf data	4
Visualization	Tools for visualizing data	1
xml test	For xml in beta version	5

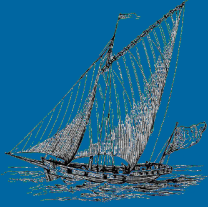
An orange abstract graphic consisting of two overlapping circles, one larger than the other, positioned to the left of the title.

PART 1

Elixir Registry



2.1 ELIXIR Tools and Data Services Registry



Why **A** registry?

A registry addresses the question of resource discovery

- Aims in **FINDING** and **UNDERSTANDING** relevant resources by various means
- Gives relevant information on that resource
- Tells how to access it (web services' url, download pages, ...)



Why **THIS** registry?

- Uses **EDAM** ontology: a **CONTROLLED** vocabulary to define bioinformatics tools operation, topics, datatypes and data formats.
- Coupled with workbenches like Moby, Galaxy, ... who have become, very useful resources providers,
 - Enabling a decentralized registry, maintained by the resource specialist themselves



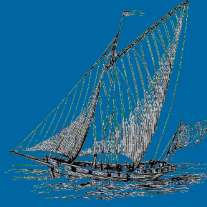


PART 2

ReGaTE



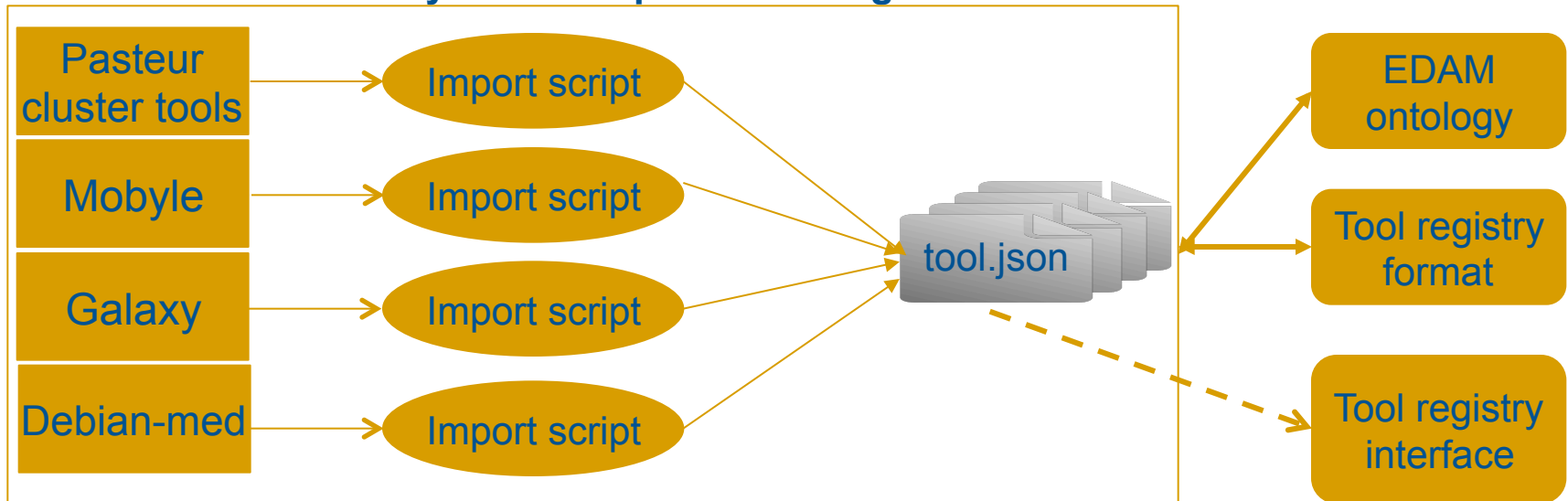
3.1 ReGaTE: History facts



• In June 2014:

- The toolinfowarehouse subproject was initiated during an EDAM meeting.

→ AIM: **key resource providers to gather data about bioinformatics tools**



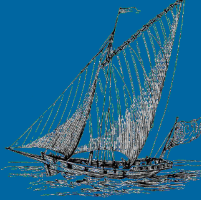
The IDEA:

→ Use them to start filling “massively” the ELIXIR Registry

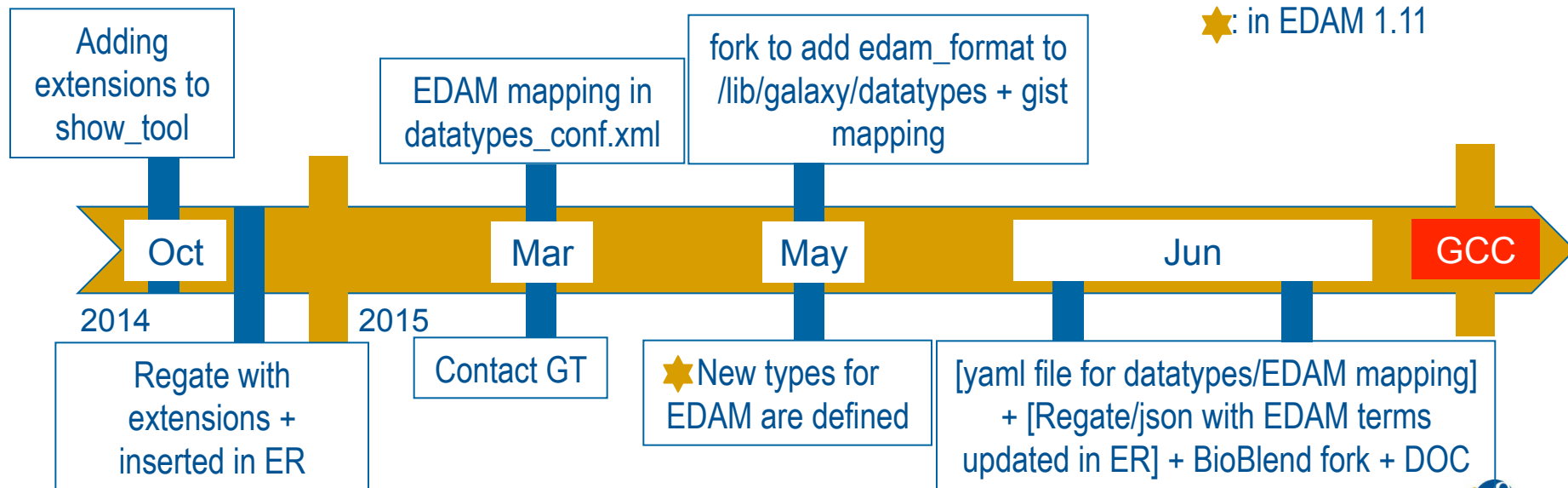
• ReGaTE, stands for Registration of Galaxy Tools in Elixir:

- Uses Bioblend API to extract information concerning installed tools on any Galaxy; tools.show_tool (io_details=True) to get input/output information
- Resulting JSON structure files for each installed tool

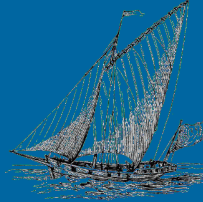
3.2 ReGaTE: Challenges



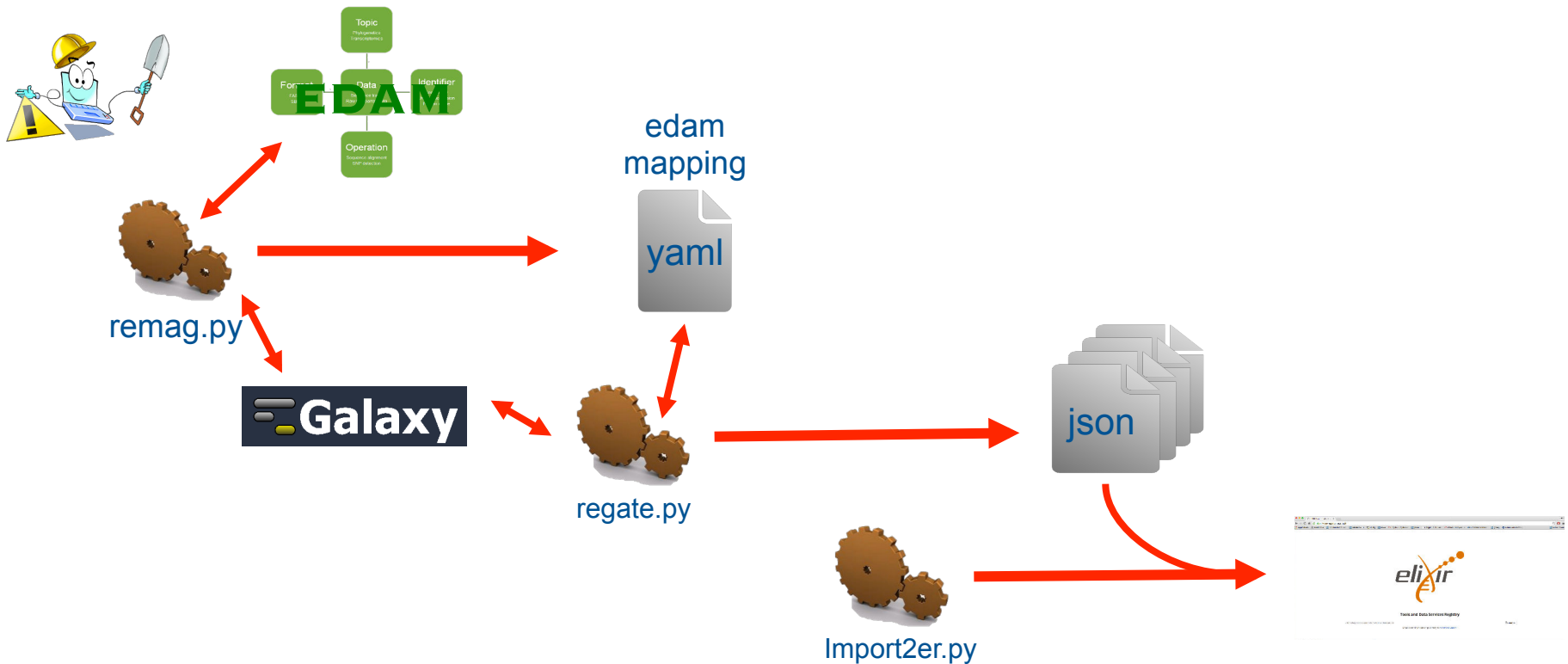
- The dictionary returned by the Bioblend `show_tool` function can be as complex as the Galaxy xml itself
- Galaxy datatypes are a challenge by themselves ☺
- The function field is a key field in Elixir Registry (ER), it gathers:
`functionName`, `functionDescription`, `functionHandle`, input [`dataFormat`, `dataType`], output [`dataFormat`, `dataType`] → the red fields are EDAM based



3.3 ReGaTE: How it works...



```
##Installation process:  
#pip install -e  
git+https://github.com/bioinfo-center-pasteur-fr/ReGaTE.git#egg=regate  
#pip install -r src/regate/requirements.txt
```



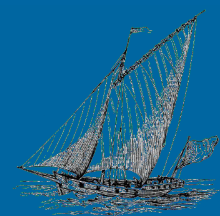


CONCLUSION

& Perspectives



To resume...

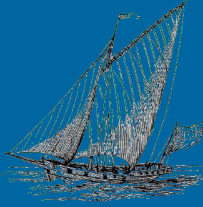


- Some of our jsons are in the registry, key words: “galaxy institut pasteur”
- If interested to display your Galaxy tools in the ER
 - Use ReGaTE to generate the json files, we tested on several Galaxy servers

Galaxy instance	Nb of jsons	connection	From where
https://galaxyapi.web.pasteur.fr	219	API key	From outside
http://usegalaxy.org	433	API Key	from outside
http://galaxy.sb-roscoff.fr/	377	API Key	ssh on roscoff
http://galaxymetabolomics4api.sb-roscoff.fr/	62	API Key	ssh on roscoff
Your server ?	?	?	?

- You only need a galaxy updated after Oct 2014, an activated API and an API key to enable the use of BIOBLEND
- To insert the JSONs in the Registry, you need to contact the registry group using the mailing list registry@elixir-dk.org.

Conclusion & Perspectives



- The ReGaTE tool is the result of a highly productive collaboration between CBS in Denmark, University of Bergen in Norway, The Institut Pasteur in Paris and the Galaxy Team.
- It is on GitHub: <https://github.com/bioinfo-center-pasteur-fr/ReGaTE>, with a first version documentation.
- The addition of EDAM in Galaxy will simplify datatypes management.
- Nice to interact with the Galaxy Team.
- To finish, I'd say that EDAM for galaxy datatypes is only a first step. Indeed, with the French Galaxy workgroup we are planning to work on an “edamization” of the toolshed. (EDAM Operations and Topics to annotate tools)



THANK YOU !

