Opportunities and needs within the consortium

Anne-Françoise Adam-Blondon, Mario Pezzotti, Holtgraewe Daniela, Paul Kersey

▶ To cite this version:

Anne-Françoise Adam-Blondon, Mario Pezzotti, Holtgraewe Daniela, Paul Kersey. Opportunities and needs within the consortium. Integrape 2019 - Data Integration as a key step for future grapevine research, Adam-Blondon A-F; Pezzotti M, Mar 2019, Chania, Greece. hal-03316648

HAL Id: hal-03316648
https://hal.inrae.fr/hal-03316648v1

Submitted on 6 Aug 2021

Data integration to maximise the power of omics for grapevine improvement

Opportunities and needs within the consortium

A-F Adam-Blondon (INRA, FR)
M. Pezzotti (U. Verona, IT)
D. Holtgraewe (U. Bielefeld, DE)
P. Kersey (Kew Garden, UK)

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

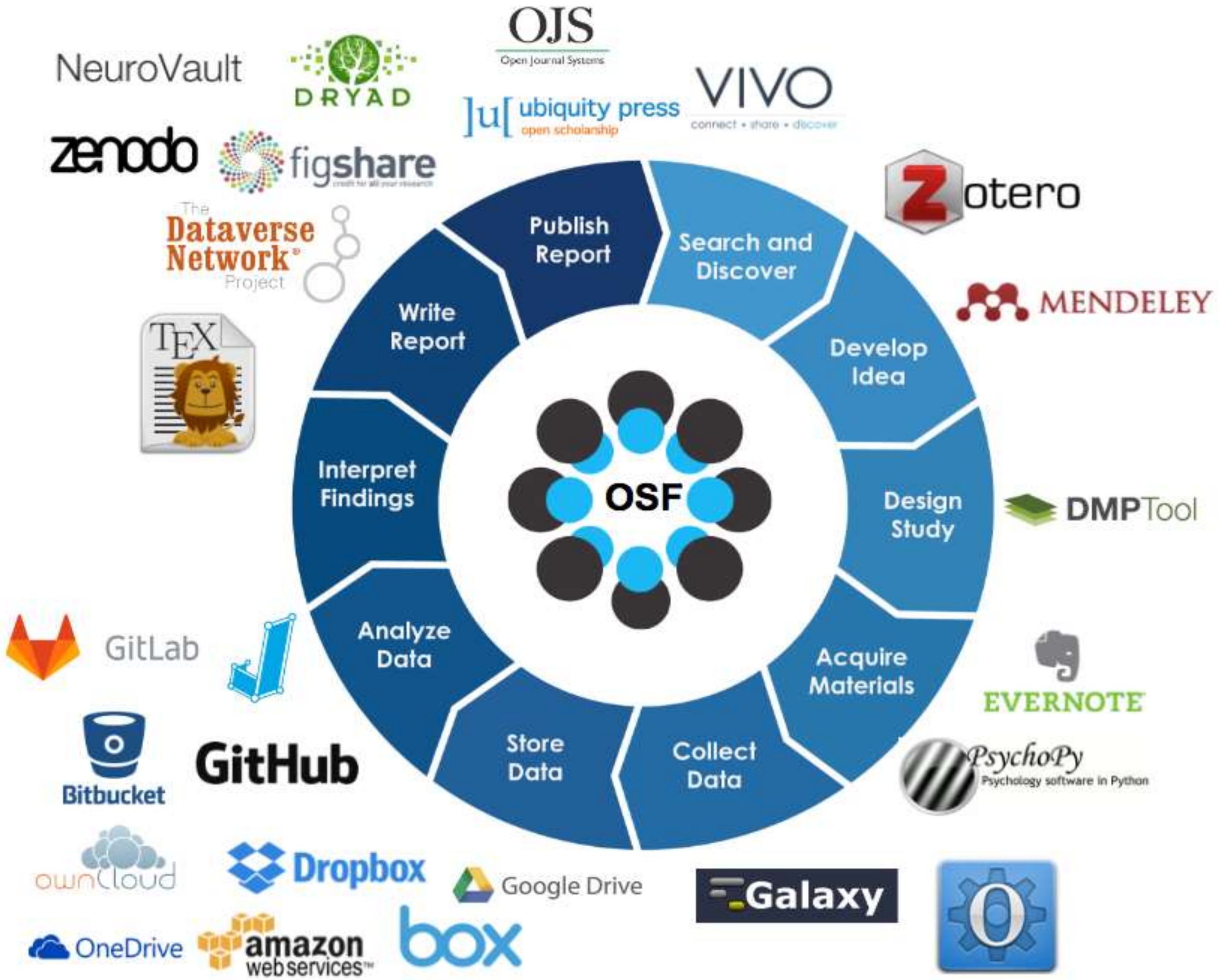
COST Action CA17111 INTEGRAPE

Europe has been continuously reinforcing its policy open access to FAIR research data

“Facilitating access to results encourages the re-use of research outputs and supports Open Science. This is essential for Europe's ability to enhance its economic performance and improve the capacity to compete through knowledge. […] Results of publicly-funded research can therefore be disseminated more broadly and faster, to the benefit of researchers, innovative industry and citizens.

Recently funded projects were asked to add a WP supporting a FAIR compliant Data Management Plan

Integrape is about facilitating FAIR data management along all its life cycle



…which is still challenging!

From Rowan University, NJ, USA

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

COST Action
CA17111
INTEGRAPE

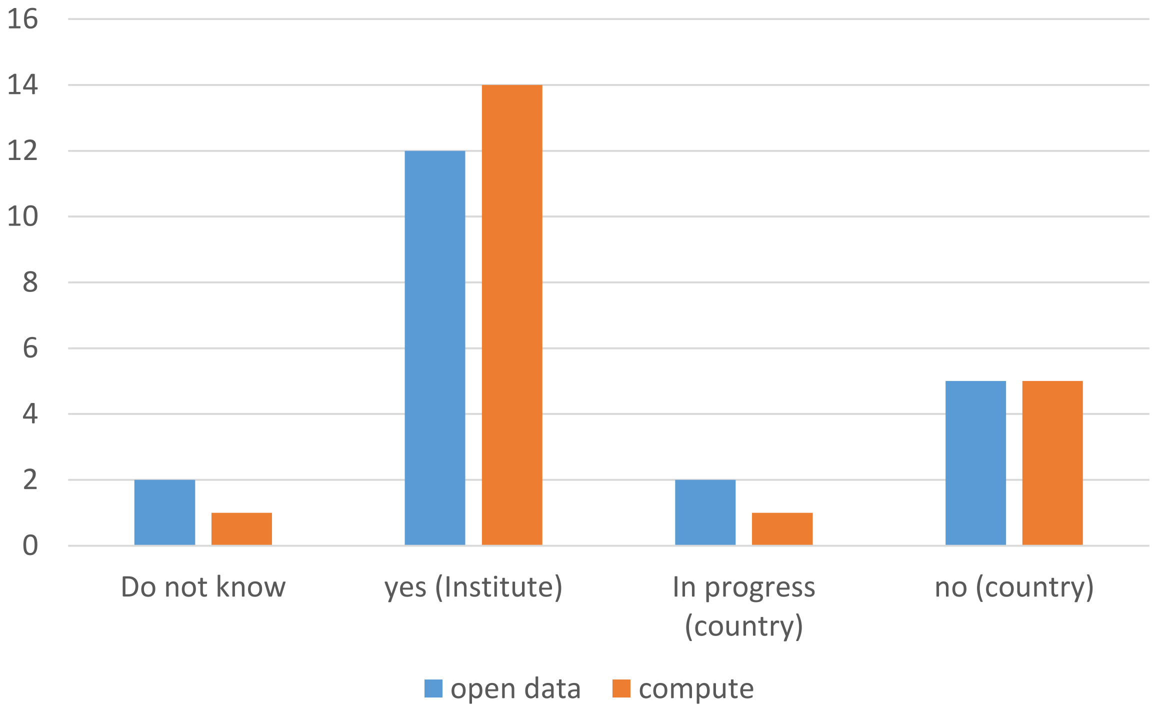
Short Survey sent to all the MC members

First part aimed at making an overview:

- National policies about open data in research
- Access to compute infrastructures
- Involvement in international infrastructure on bioinformatics or open data
- Use of database for the storage of data produced by the research units
- Needs in terms of training in bioinformatics
- Needs in terms of infrastructures

➡ Answers from 21 institutes of 16 countries

Country/Institute policies in terms of open data and of access to computing resources



10 institutes/countries involved in ELIXIR infrastructure, 2 in Life Watch +/- GBIF, 1 in DISSCO, 1 in PRACE, 1 in EMBL

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
Funded by the Horizon 2020 Framework Programme of the European Union
COST Action CA17111 INTEGRAPE

What database/ computing infrastructure do you use?

Database

3 -> no use

15 -> local database maintained most of the time internally by its creator, sometimes with the help of IT central services

Commons -> international data repositories: VIVC, vitis-eu, INSDC archives (Genbank and co)

Computing infrastructure

11 -> no use

8 -> internally maintained servers

5 -> Access to (high performance) clusters maintained by the institute or at the national level

2 -> Access to national super computing facilities

1 -> commercial cloud

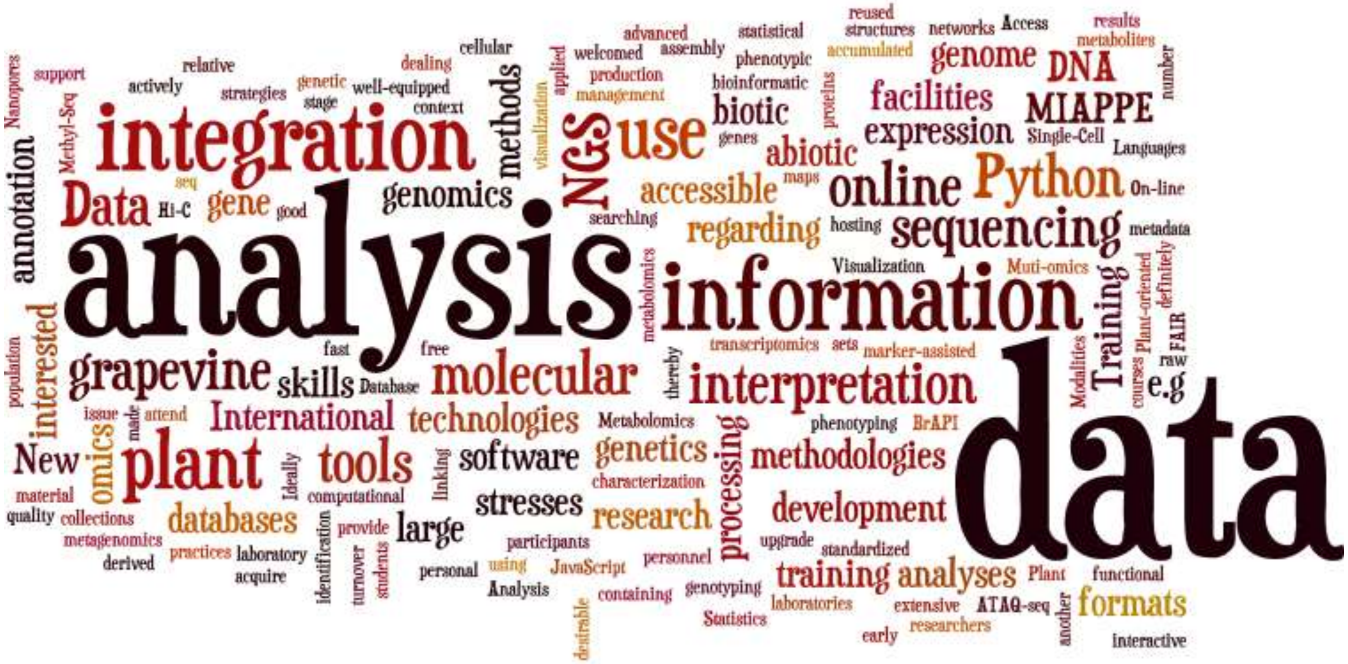
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

COST Action CA17111 INTEGRAPE

What would be the priority in terms of training?

- Genomic data analysis is the first need and to a lesser extent data standardization
- The target is mostly students but some mentioned also the management
- Online material, resources to allows students to attend courses are problematic for some respondents



COST
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

COST Action
CA17111
INTEGRAPE

What would be the priority in terms of access to a bioinformatic infrastructure?

- Access to stable/reliable tools for omics data analysisis the most frequently cited need
- Storage capacity
- Help for the development of a policy and legislation compliant infrastructures
- National and transnational access (without the need to develop ad hoc collaborations) to powerful computing infrastructures

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

COST Action CA17111 INTEGRAPE

Short Survey sent to all the MC members

The second part aimed at digging a bit more into the data you produce and the way you manage it:

- What types/volumes of genomics-related data are you currently producing?
- How do you expect this to change over the next 5 years?
- How do you currently publish the data? Do you publish them at all?
- What are the obstacles to publishing the data?
- Are there data types you cannot publish?
- When you seek to use relevant published data, is it FAIR (Findable, Accessible, Interoperable, Re-usable)? If not, why?
- Are there existing norms in the area of data publishing that you adhere to?
- Are there candidate norms that you would like the community to adopt?
- Are there new norms that the community could usefully develop?
- Does the grape research community need its own dedicated Grape Information System, and if so, what would you like to be able to use or to contribute?

→ Answers from 10 institutes

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme
of the European Union

COST Action
CA17111
INTEGRAPE

Type and volume of data produced, evolution in the next 5 years

- Metabolomics data/ a few hundreds of genotypes
- Whole genomes assemblies and annotations of V. vinifera and non-vinifera species; Annotations of specific gene families or genomic features (e.g. resistance genes, TE)
- Hundreds of RNAseq datasets every year
- Markers and genotyping data (SSR, SNP, SRAP, …) from a large range of techniques; Maps, QTL,…
- Polymorphism, re sequencing data, RAD seq
- Phenotyping data
- Genetic resources
- Simulated data

Evolution

- Increase of the pace and volume of data
- New opportunities linked to technical changes but raising questions about references, standards, …: pangenomics, single cell RNAseq, RiboSeq, …

Funded by the Horizon 2020 Framework Programme
of the European Union

Practices in terms of data publication; possible bottlenecks

Most partners publish their data in international archives or file repositories

Bottlenecks

- **Mostly the peer review process of paper associated to the data**
- Lack of standards for some datatypes
- Workflow issue / Lack of dedicated and skilled persons
- Nagoya protocol

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme
of the European Union

COST Action
CA17111
INTEGRAPE

Can you effectively re-use already published data?

Many issues mentioned:

- Gene ID , identification of the plant material unclear or obsolete
- Incomplete metadata or in a challenging format for re-use
- Incomplete data (e.g. primer sequence) or low quality data

Whishes for the future

- **Open access to data, better published along with papers**
- Need for the development of non existing standards: e.g. in relation with pan genomics, API for metabolomics, …
- A hub for grapevine research is definitely welcome

EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme
of the European Union

COST Action
CA17111
INTEGRAPE