Survey of information systems and web services useful for the grapevine community - Gaps and Recommendations

Date : 2022 September 26th

Authors and affiliation

Author	Affiliation	ORCID
Anne-Françoise ADAM-BLONDON	Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France	0000-0002-3412-9086
Michael ALAUX	Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France	0000-0001-9356-4072
Jérôme GRIMPLET	CITA-Aragon. Plant science department. 50059 Zaragoza, Spain	0000-0002-3265-4012
Daniela HOLTGRAEWE	Bielefeld University, Chair of Genetics and Genomics of Plants, Faculty of Biology & Center for Biotechnology (CeBiTec), Bielefeld, Germany	0000-0002-1062-4576
Paul KERSEY	Royal Botanic Garden, Kew, UK	0000-0002-7054-800X
José Tomas MATUS	Institute for Integrative Systems Biology, I²SysBio (Universitat de València - CSIC), 46908, Paterna, Valencia, Spain.	0000-0002-9196-1813
Marco MORETTO	Fondazione Edmund Mach, Research and Innovation Centre, via E. Mach, 1, 38098 San Michele all'Adige (TN), Italy	0000-0003-4555-7243
Mario PEZZOTTI	Department of Biotechnology,	0000-0002-7430-6147

	University of Verona, Strada Le Grazie 15, 37134, Verona, Italy	
Stefania PILATI	Fondazione Edmund Mach, Research and Innovation Centre, via E. Mach, 1, 38098 San Michele all'Adige (TN), Italy	0000-0001-7103-9613
Camille RUSTENHOLZ	Université de Strasbourg, INRAE, SVQV, 68000 Colmar, France	0000-0001-5355-3408
Reinhard TÖPFER	Julius Kühn-Institute (JKI), Institute for Grapevine Breeding Geilweilerhof, 76833 Siebeldingen, Germany	0000-0003-1569-2495
Amandine VELT	Université de Strasbourg, INRAE, SVQV, 68000 Colmar, France	0000-0003-2368-839X

Acknowledgements

This deliverable was developed upon work from COST Action CA 17111 INTEGRAPE, supported by COST (European Cooperation in Science and Technology).

Content

Context and objectives

The international grapevine community has been working for a decade on recommendations and best practices to facilitate data exchange and reuse. Its first activity consisted in the development of a nomenclature system for grapevine genes that would allow consistent gene naming[1]. The second was to propose a strategy aiming at facilitation grapevine data findability and reuse through recommendation on metadata standards, identifiers associated to samples which use will in turn facilitate the development of a centralized search tool across a federation of databases[2]. The aim of the network was to make a step forward towards the implementation of this strategy by working on the three complementary pillars of success:

- improvement of the standards, ontologies and of guidelines for FAIR data management necessary to the grapevine community
- dissemination and training on these guidelines
- development of a federation of databases holding grapevine data:
    - sharing the same metadata standards and identifiers
    - being all searchable through the same central portal
    - allowing data access and reuse

The first objective of this document is to set up the scene of the third pillar described above by:

- identifying the databases/information systems frequently used by the grapevine community
- checking how data (including metadata) can be accessed from these information systems
- Identifying possible problems and gaps

The second objective is to make recommendations on steps to move forward towards the building of a federation of databases that would collaborate to provide central data services to the grapevine research community[3].

List of Information Systems, databases and datasets used/developed by the grapevine community that networked under the auspices of the Integrape COST action

A list of databases has been developed from three sources:

---

[1] Grimplet J, Adam-Blondon A-F, Bert P-F, Bitz O, Cantu D, Davies C, Delrot S, Pezzotti M, Rombauts S, Grant R Cramer GR (2014) The grapevine gene nomenclature system. BMC Genomics, 15 :1077 (https://doi.org/10.1186/1471-2164-15-1077). See also the digest on the web site of the Integarpe COST action: https://integrape.eu/resources/genes-genomes/reference-gene-catalogue-and-nomenclature-recommendations/

[2] A-F Adam-Blondon, M Alaux, C Pommier, D Cantu, Z-M Cheng, GR Cramer, C Davies, S Delrot, L Deluc, G Di Gaspero, J Grimplet, A Fennell, JP Londo, P Kersey, F Mattivi, S Naithani, P Neveu, M Nikolski, M Pezzotti, BI Reisch, R Töpfer, MA Vivier, D Ware, H Quesneville (2016) Towards an open grapevine information system. Hort Res, 3, 16056. doi:10.1038/hortres.2016.56. The approach described in this paper has been implemented for 15 years by the WheatIS working group of the wheat community (http://wheatis.org/) and its central portal allowing to search data across a federation of information systems scattered across the world (https://urgi.versailles.inrae.fr/wheatis/)

[3] as it is proposed in the GRAPEDIA project funded by the COST association: https://integrape.eu/wp-content/uploads/2022/09/CIG-Application-GRAPEDIA_CA17111_2022.pdf

- A survey on the current practices and needs in terms of bioinformatics of the COST participants about the data resources they use the most (Annexe 1).
- A request to the network to complete the list with some softwares or databases they would be aware of and that they would think of interest
- A quick review of the recent literature with the keywords ("database", "grapevine" and "information system" and by checking the status of the content of a 10 years old review on bioinformatic tools for the grapevine community[4].

The resulting list of 28 databases is presented in table 1. Ten repositories are generic in terms of organisms: the category regroups the international genomic archives developed by the International Nucleotide Sequence Data Consortium (INSDC; EMBL-EBI, NCBI, Kyoto Genomic Institute) and the dataverse of the Portuguese node of the ELIXIR ESFRI. Eight repositories are holding data from different plants with very different scopes: from repositories that can handle different types of data (genetic resources, phenotypes, various genomic data) with variable level of curation among the datasets to others that are very focussed on one type of genomic data (e.g. lncRNA annotations), usually more uniform in terms of curation (but not always). Finally, 10 repositories are holding only grapevine highly curated data, usually focussing on one type of data (genetic resources, transcription data, etc…).

Table 1 also presents the possibilities offered by the databases in terms of access to data through different mechanisms: FTP, after filtering and queries, through API, etc… Finally, a link to the description of the repository in the FAIRsharing[5] or in the Re3Data[6] registries is given as a source of additional information on its maintenance status, links to other databases and use of standards.

[4] Grimplet J, Dickerson J, Adam-Blondon A-F, Cramer G (2011) Bioinformatics Tools in Grapevine Genomics. *in:* A-F Adam-Blondon, JM Martinez-Zapater, Chittaranjan Kole (eds) *Genetics, Genomics and Breeding of Grapes.* Science Publishers and CRC Press. pp 317-331 (file:///C:/Users/urgi-159/Downloads/10.1201_b10948_previewpdf-1.pdf)

[5] FAIRsharing (https://fairsharing.org/) is a curated catalog of data and metadata standards, inter-related to databases and data policies. It affects to each described resource a unique identifier (recently moved into a DOI) and gives some insights on the level of maintenance and of interoperability of the resources. It also allows the description of collections of resources developed by a given community to support FAIR data.

[6] re3data (https://www.re3data.org/) is a curated registry of data repositories available to the research community. re3data is a DataCite partner service (https://datacite.org/partnerservices.html) with an international editorial board that spans the globe and covers every research domain. It has collaborations with FAIRsharing.

Table 1. List of databases used by the grapevine community.

Name	Website	Short description	Grapevine data content	Modalities for accessing data	Restrictions to data access?	Data visualization and/or analysis?	Type of data	Primary Archive /added value data set	FAIR/Open Science compliance as reviewed in the FAIRsharing portal or the Registry of Research Data Repositories
ArrayExpress	https://www.ebi.ac.uk/arrayexpress/	ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.	131 experiments/datasets tagged "Vitis" (August 2022)	*- Bulk and custom data access* *- In-house RESTful APIs*	NO	YES	Generic	Added value	https://fairsharing.org/FAIRsharing.6k0kwd
Biodata.pt dataverse	https://dmportal.biodata.pt/	ELIXIR-PT portal of datasets which is part of a set of bioinformatic services to the Portuguese life science community (see https://biodata.pt)	One grapevine dataset found so far	*- downloads* *- in house and BrAPI compliant RESTful APIs*	NO	NO	Generic	Primary	NONE
CANTATAdb 2.0	http://rhesus.amu.edu.pl/CANTATA/	CANTATAdb 2.0 is a database of lncRNAs identified computationally in 36 plant species and 3 algae, Vitis vinifera included	Grapevine reference genome sequence annotated	- downloads	SOMETIMES	YES	Plant	Added value	NONE
ENA	http://www.ebi.ac.uk/ena/	The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.	A few genome assemblies of different grapevine accessions, expression atlas, pathogen sequences, etc...	*- Bulk and custom data access* *- In-house RESTful APIs*	NO	NO	Generic	Added value	https://fairsharing.org/FAIRsharing.dj8nt8
Ensembl Plants	https://plants.ensembl.org/index.html	Plant comparative genomic portal of the European Bioinformatic Intistute (EMBL-EBI	- Grapevine genome assemblies used: V. vinifera cv PN40024 v1 - Contains SNP data from Myles et al (2010) 10.1371/journal.pone.0008219, probe data from the GeneChip™ Vitis vinifera Genome Array, (https://www.thermofisher.co	*- Bulk and custom data access* *- In-house RESTful and Perl APIs*	NO	YES	Plant	Added value	https://fairsharing.org/FAIRsharing.j8g2cv

			m/order/catalog/product/900509#/900509) - Comparative genomic analysis with 92 other plant species.						
EVA	http://www.ebi.ac.uk/eva/	The European Variation Archive is an open-access database of all types of genetic variation data from all species.	Only an exemplar data set on grapevine	*- Bulk and custom data access* *- In-house RESTful APIs*	NO	YES	Generic	Added value	https://fairsharing.org/FAIRsharing.6824pv
GenBank	https://www.ncbi.nlm.nih.gov/genbank/		6323 datasets using samples from the Vitis genus	*- Bulk and custom data access* *- In-house RESTful APIs*	NO	NO	Generic	Added value	https://fairsharing.org/FAIRsharing.9kahy4
Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/	GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted.	Over 5,500 datasets from the Vitis genus (August 2022)	*- Bulk and custom data access* *- In-house RESTful and Perl APIs*	NO	NO	Generic	Added value	https://fairsharing.org/FAIRsharing.5hc8vt
GnpIS	https://urgi.versailles.inrae.fr/gnpis	INRAE, curated and integrated data resources for plant genetics and genomics	- URGI Vitis vinifera JBrowse (https://urgi.versailles.inrae.fr/jbrowse/gmod_jbrowse/?data=myData/Vitis/data_gff) with the current version of the PN40024 reference genome assembly with all the gene annotation versions. - Passport data of the french Grapevine germplasm collection and some phenotypic data: https://urgi.versailles.inrae.fr/faidare/?sources=https:%2F%2Furgi.versailles.inra.fr%2Fgnpis - Genotyping data: https://urgi.versailles.inrae.fr/GnpSNP/snp/genotyping/form.do#results - genetic maps: https://urgi.versailles.inrae.fr/GnpMap/mapping/searchMap.do	*- downloads* *- in house and BrAPI compliant RESTful APIs*	SOMETIMES	YES	Plant	Added value	https://fairsharing.org/FAIRsharing.dw22y3
GoMapMan	http://www.gomapman.org/	GoMapMan is an open web-accessible resource for gene functional annotations in the	Grapevine data available	*Download*	NO	YES	Plant	Added value	https://fairsharing.org/FAIRsharing.9ry4cz

		plant sciences. It was developed to facilitate improvement, consolidation and visualization of gene annotations across several plant species. GoMapMan is based on the plant specific MapMan ontology (link is external), organized in the form of a hierarchical tree of biological concepts, which describe gene functions.							
Gramene	http://www.gramene.org/	Curated, open-source, integrated data resource for comparative functional genomics in crops and model plant species, including Vitis vinifera.	1,461,370 entries for the Vitis genus	*- Bulk and custom data access* *- In-house RESTful APIs*	NO	NO	Plant	Added value	https://fairsharing.org/FAIRsharing.zjdfxz
Grape eFP Browser	http://bar.utoronto.ca/efp_grape/cgi-bin/efpWeb.cgi	This tool exploring the transcriptional atlas of Vitis vinifera (cv. Corvina).	The atlas consists in 54 grapevine samples (including different developmental stages of bud, inflorescence, tendril, leaf, stem, root, developing berry, withering berries, seed, rachis, anther, carpel, petal, pollen, and seedling) with three biological replicates for each sample. Fasoli et al. (2012). https://doi.org/10.1105/tpc.112.100230	*Download*	NO	YES	Grapevine	Added value	NONE
GRape Expression ATlas (GREAT)	https://great.colmar.inrae.fr/	GREAT allows to analyze and visualize public RNA-seq data from Vitis vinifera species. The application performs also some statistical analyses, like genes clustering with MFuzz or differential gene expression analysis with askoR (edgeR).	The application is based on more than 2600 public RNAseq samples of Vitis vinifera available in the SRA database. For the moment the data is analyzed on PN12X.V2 ; integration of the analyzed data on PN40024.v4 is in progress.	*Download*	YES	YES	Grapevine	Added value	NONE
GrapeGenomics	http://www.grapegenomics.com/	A web portal with genomic data and analysis tools for wild and cultivated grapevines	Wild and cultivated Vitis genomes. BLAST. Network analysis. Genome browser.	- GUI - Download	SOMETIMES	YES	Grapevine	Primary	NONE
Grape RNA	http://www.grapeworld.cn/gt/index.php	Grape RNA is one part of the grape world. The main work of this database is to collect,		*Download*	NO	YES	Grapevine	Added value	NONE

		store, treatment and share the Transcriptome data of the vitis genus.						
Kyoto Encyclopedia of Genes and Genomes (KEGG	https://www.kegg.jp/kegg/	KEGG is a database resource for understanding high-level functions of biological systems from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.	https://www.kegg.jp/kegg-bin/show_organism?category=Plants	*- Bulk and custom data access* *- In-house RESTful APIs*	SOMETIMES	YES	Generic	Added value	https://fairsharing.org/FAIRsharing.327nbg
MetaboLights	https://www.ebi.ac.uk/metabolights/	MetaboLights is a database for metabolomics studies, their raw experimental data and associated metadata. The database is cross-species and cross-technique and it covers metabolite structures and their reference spectra as well as their biological roles and locations.	23 studies and 313 coumpounds tagged Vitis vinifera (august 2022)	*- Bulk and custom data access* *- In-house RESTful APIs*	NO	YES	Generic	Added value	https://fairsharing.org/FAIRsharing.kkdpxe
OneGenE_Vitis	vitis.onegenexp.eu	OneGenE explores causal relationship among grapevine genes. It computes ranked candidate gene lists causally related to an input gene using the expression data stored in Vespucci.	http://ibdm.disi.unitn.it/onegene/vv/onegene-vv.php	*Download*	NO	NO	Grapevine	Added value	NONE
Phytozome	https://phytozome-next.jgi.doe.gov/	Plant Comparative Genomics portal of the Department of Energy's Joint Genome Institute	12X March 2010 release of the draft genome and v2.1 annotation of Vitis vinifera by the French-Italian Public Consortium for Grapevine Genome Characterization	*- Bulk and custom data access* *- Globus access* *- RESTful APIs*	NO	YES	Plant	Added value	https://fairsharing.org/FAIRsharing.83d06b
Plant Metabolic Networks (PMN) databases	https://plantcyc.org/	The PMN currently houses one multi-species reference database called PlantCyc and 126 species/taxon-specific databases.	https://pmn.plantcyc.org/organism-summary?object=GRAPE	*- downloads (smart tables* *- RESTful APIs*	NO	YES	Plant	Added value	NONE
PlantRegMap	http://plantregmap.gao-lab.org/	Provides a comprehensive, high-quality resource of plant transcription factors (TFs), regulatory elements and	Some datasets are available for Vitis vinifera	*Download*	NO	NO	Plant	Added value	NONE

		interactions between them, advancing the understanding of plant transcriptional regulatory system.						
PRIDE	https://www.ebi.ac.uk/pride/	The PRIDE PRoteomics IDEntifications (PRIDE) Archive database is a centralized, standards compliant, public data repository for mass spectrometry proteomics data, including protein and peptide identifications and the corresponding expression values, post-translational modifications and supporting mass spectra evidence (both as raw data and peak list files)	32 grapevine tagged datasets (august 2022)	*- Bulk and custom data access* *- In-house RESTful APIs*	NO	YES	Generic	Added value	https://fairsharing.org/FAIRsharing.e1byny
SRA	https://www.ncbi.nlm.nih.gov/sra	Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data.	https://www.ncbi.nlm.nih.gov/sra/?term=vitis%5BOrganism%5D	*- Bulk and custom data access* *- In-house RESTful APIs*	NO	NO	Generic	Primary	https://fairsharing.org/FAIRsharing.g7t2hv
The European Vitis Database	http://www.eu-vitis.de	The European Vitis Database gathers passports and descriptors about the grapevine accessions maintained by the european genbanks of the European Cooperative Programme for Plant Genetic Resources (ECPGR; https://www.ecpgr.cgiar.org/)	The most comprehensive catalog of grapevine accessions in European genbanks. Perhaps even better referenced than Eurisco (the ECPGR genbank catalog).		SOMETIMES	NO	Grapevine	Added value	https://www.re3data.org/repository/r3d100013219
VESPUCCI	http://vespucci.fmach.it	VESPUCCI is a comprehensive grapewine cross-platform gene expression database. The compendium was carefully constructed by collecting, homogenizing and formally annotating publicly available microarray and RNA-seq experiments from Gene Expression Omnibus (GEO),	VESPUCCI contains 3682 microarray and 3598 RNA-seq samples measured across 271 experiments all formally annotated using bio ontologies. Data are normalized using TPM (transcript per million) and LogRatios. The tools provided allow to explore patterns of gene expression	*- Download of the whole compendium* *- Programmatic access through a GraphQL API* *- Programmatic access through Python and R packages*	NO	YES	Grapevine	Added value	NONE

		Sequence Reads Archive (SRA) and ArrayExpress.	across different conditions and build heatmaps and co-expression networks.					
Vitis International Variety Catalogue (VIVC)	https://vivc.de	VIVC is an encyclopedic database with around 23000 cultivars, breeding lines and Vitis species, existing in grapevine repositories and/or described in bibliography. It is an information source for breeders, researchers, curators of germplasm repositories and interested wine enthusiasts. Besides cultivar specific passport data, SSR-marker data, comprehensive bibliography and photos are to be found.	The most comprehensive and well curated knowledge base on grapevine varieties. The reference database for varieties identifiers, crucial to distangle the fuzzy naming of varieties.	- *downloads*	NO	NO	Grapevine	Added value	https://www.re3data.org/repository/r3d100013221
VitisNet	https://github.com/jgrimplet/VitisNet	A web application based on the open source software Cytoscape (www.cytoscape.org) that allows to browse networks of gene regulation evidenced by a "dormancy" data set.	https://www.sdstate.edu/agronomy-horticulture-plant-science/functional-genomics-bud-endodormancy-induction-grapevines	- downloads	SOMETIMES	YES	Grapevine	Added value	NONE
Vitis Visualization Platform (VITVIZ)	http://www.vitviz.tk/	Web-based platform with several tools to visualize gene-related data for gene lists: gene expression data, and gene co-expression networks, JBrowser with DAP-seq published data, search engine for looking correspondencies between all PN40024 genome annotations and the Grape Reference Catalogue with interactive view of global gene expression in the SRA database.	(1) Grape Gene Reference Catalogue; (2) PN40024 ID Conversion tool; (3) Genome Browser for DAP-seq data (DAP-Browse); (4) Expression atlases (Corvina developmental atlas, Pinot flower atlas, Botrytis atlas, Hydric stress atlas, Cabernet/Merlot Berry development atlas); (5) EXHARA (expression visualization in all SRA public data); (6) TRANSMETAdb (gene-to-metabolite correlation tool); (7)AggGCN (aggregated gene co-expression network visualizator).	- *downloads*	NO	YES	Grapevine	Added value	NONE

Problems and gaps

In terms of databases

The list is probably not comprehensive and skewed to the community that was active in the working groups 1 and 2 of the action. The vast majority of the databases presented in table 1 are targeted to genomic data. The exceptions are GnpIS and BioData.pt that allow the storage of other types of data (genetic resources and phenotyping for GnpIS and any type of data for BioData.pt) and of course, The European Vitis Database and the Vitis International Variety Catalog, which are dedicated to grapevine genetic resources. If the centralization of data on genetic resources and of genomic data is continuously improved under the auspices of international bodies such as respectively the Food and Agriculture Organization (FAO) and the INSDC, phenotyping data is still lacking a global portal and many countries do not propose central archives to their research community.

Community databases often prove to be very unstable over time, with many links to the resources that do not pass after 10 years or evidence for non maintenance (no update of the data, no evolution of the interface). The only exceptions are the two databases in relation with grapevine genetic resources, The European Vitis Database and the Vitis International Variety Catalog that have been maintained and curated by JKI (Germany) for several decades now. They are also the most popular databases in the survey made at the beginning of the project followed by the resources developed by the NCBI and EMBL-EBI (Annexe 1). This emphasizes the importance of long term sustainability for adoption by users and for data reuse. Many of these community databases also have similar objectives (e.g. develop compendium of curated transcriptome data, curate metabolic networks) and it is often difficult to see whether their data aggregate *in fine* somewhere (e.g. in reactome, BioCyc, Gramene/ENSEMBL Plants, …). Even when not maintained anymore, these databases are still representing an interesting source of curated data sets that could be better managed by the community.

Data findability is difficult due to the multiplicity of sources, in particular when data is not available in international archives with well maintained central portals enabling high data findability.

In terms of data access

Access to data is highly heterogeneous and often poorly documented in terms of accessed file formats in the community databases (information can be difficult to find). Most of the time, in these databases, data access is possible via downloads and is not machine actionable through API/web services. Opposite, generic databases provide diverse types of access to data, including via API/web services.

Nearly no databases, generic or community driven, has implemented the standard API dedicated to plant genetic and genomic data for breeding applications, BrAPI[7]. This might reflect a need to quantitatively increase the informatic skills of the plant community to allow reaching a step that is a bit labor intensive at first and a lack of current alignment of BrAPI with the API developed by generic repositories.

[7] Selby et al, BrAPI an application programming interface for plant breeding applications, *Bioinformatics*, Volume 35, Issue 20, 15 October 2019, Pages 4147–4155, https://doi.org/10.1093/bioinformatics/btz190

Recommendations to enhance Open Data for grapevine research

Use as much as possible generic/Plant dedicated databases and associated resources (message to the PIs)

Building on already existing databases will allow globally the grapevine community to put its resources in:

- the curation, standardization and publication of grapevine datasets and the development of associated specific standard vocabularies, ontology terms, protocols, scales (e.g. developmental stages), …
- the development of data management strategy including long term archival of these data sets
- added value software/infrastructure development aiming at:
    - the implementation of local data management workflows facilitating FAIR data management
    - contributing to the improvement of generic/plant databases/resources
    - providing applications using the APIs of the generic/plant databases and providing added value through outputs in standard formats (interoperable with the rest of the ecosystem).
- disseminating good practices and training the community.

In turn, the Grapevine community will benefit from:
- a better findability of its data, less scattered in databases likely to be more sustainable because supported by more than one community
- a more efficient use of its resources and with more impact on the research community
- knowledge and experience from other communities.

Engage as a group with communities actively working in the field of data management, data curation and integration (message to data managers and developers)

The key point is to develop a grapevine international working group for grapevine FAIR data with all the skills necessary (project management, dissemination, data curation, software development, ontology, …) to prioritize and coordinate its activities and develop collaborations with other groups. The Integrape COST action has been one opportunity and the Grapedia project will be the next one. However, projects might structurally generate skewed working groups and/or misalignment between the priorities of the global grapevine community and the output of the project: there is a need to be aware of this potential problem and to take measures to avoid it (e.g. through a scientific advisory board, through outreach events targeting the global community, …).
A working group of the grapevine community aiming at facilitating the management and access to FAIR grapevine data should aim at including representatives or liaisons with the most important data repositories holding grapevine data. The list of repositories shown in Table 1 is a first check list that can be exploited for that purpose. But it is to be completed with other lists set up by communities involved in plant data management as for instance:

- the ELIXIR Plant Science community[8] and its pages in the portal of resources for data management[9]
- the AgBioData international consortium[10] and its list of databases[11]
- the resources for data management highlighted by the platform for Big Data in Agriculture coordinated by the CGIAR[12]

These consortia together with others, possibly more generic in terms of organism studied (See for instance the list of ELIXIR communities at https://elixir-europe.org/communities), might also help to identify unexploited opportunities for the grapevine community.
In the end, such networking activities will help to improve and regularly update recommendations to the whole grapevine community on repositories to be used provided on community web pages[13]

Develop a culture of FAIR and TRUST

At the level of the metadata, several consortia are actively developing in collaboration with the ELIXIR and AgBioData communities a suite of metadata standards for plant data interoperability : the International Plant Phenotyping Network[14], the MIAPPE[15] and Crop Ontology initiatives. These standards are embedded in a technical standard for web services by the Breeding API[16] initiative and standard formats for genomic data are also better specified by ELIXIR and AgBioData. The output of all these activities are already being very actively taken over by the grapevine community to make their specific datasets and databases more Findable, Interoperable, Accessible and Reusable (FAIR) and therefore increase their overall impact and usefulness. The COST integrape action has contributed to develop the culture of FAIR data management in the grapevine community through training, exchanges between laboratories and conferences. The next step is to continue to sustain a working group that will continue coordination and capacity building in this area: the Grapedia project is an opportunity that will be ceased.
However, other matters than the FAIR principles are important for the long term impact of information systems. This has been first addressed through the development of certifications systems such as for instance the CoreTrustSeal[17]. Recently these various initiatives and their associated criteria of trustworthiness were re-visited to develop simple guidelines adapted to a large audience, the TRUST principles[18]. TRUST stands for Transparency (to be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence), Responsibility (to be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service), User focus (to ensure that the data management norms and expectations of target user communities are met), Sustainability (to sustain services and preserve data holdings for the long-term) and Technology (to provide

---

[8] https://elixir-europe.org/communities/plant-sciences

[9] RDMkit: https://rdmkit.elixir-europe.org/plant_sciences

[10] https://www.agbiodata.org/

[11] https://www.agbiodata.org/databases

[12] see the communities of practices on data management and on ontologies in particular at: https://bigdata.cgiar.org/communities-of-practice/

[13] e.g. the “Resource” page of the Integrape web site: https://integrape.eu/resources/data-management/

[14] IPPN: https://www.plant-phenotyping.org/

[15] www.miappe.org

[16] BrAPI: www.brapi.org

[17] https://www.coretrustseal.org/

[18] Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* **7**, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

infrastructure and capabilities to support secure, persistent, and reliable services)[14]. In the domain of the Life Sciences, these criteria are grounding the selection of sets of labeled databases that need to be sustained with transnational funding mechanisms: the ELIXIR Core Data Resources[19] and Deposition Databases for Biomolecular data[20] and since 2022, the Global Core Biodata Resources[21] of the Global Biodata Coalition[22]. These three labels include in addition several criteria in relation with the scientific impact of the database and aim at strengthening the long term support of the labeled resources. Encouraging the databases maintained by the grapevine community to self-evaluate against the TRUST principles would probably lead to a global improvement of their technical quality, of their governance and of their focus in complement to the international archives that are likely to be progressively labeled by the Global BioData Coalition. It would strengthen the quality of a federation of grapevine databases and the interest for the research community of a common data portal.

Disseminate, train and build capacity

A coordinated activity aiming at developing a federation of grapevine data would benefit from annual dissemination events using for instance the two series of conferences of the grapevine community, the Grapevine Physiology and Biotechnology conferences and the Grapevine Genetic and Breeding conferences. This would enhance adoption and engagement with the strategy of the coordinated activity and give regular feedback to the databases/portals maintainers on users expectations.

Capacity building within the community is also crucial for adoption and engagement. Common and reusable training material could be developed by an *ad hoc* working group, possibly with other initiatives having common training goals.

Conclusions

The activities carried out during the COST action INTEGRAPE allowed to develop a vision of the current landscape of databases and repositories used by the grapevine research community. Together with the other outputs of the action (Guidelines for FAIR data management, new version of the grapevine reference genome, trainings for data analysis and FAIR data management) this vision will support the development of a central data hub for the grapevine research community, which is still an objective to be achieved. This is the goal of the GRAPEDIA project starting In 2022.

[19] https://elixir-europe.org/platforms/data/core-data-resources

[20] https://elixir-europe.org/platforms/data/elixir-deposition-databases

[21] Global Biodata Coalition. (2022). Global Core Biodata Resources: Concept and Selection Process. https://doi.org/10.5281/zenodo.5845116

[22] https://globalbiodata.org

Annex 1 - CA17111 – Integrape survey on the needs of the grapevine community in terms of bioinformatics

Date: 2019, April 16th

Authors

A-F Adam-Blondon (INRAE, France), D. Holtgräwe (Bielefeld University, Germany), P. Kersey (Key Garden, UK), M. Pezzotti (University of Verona, Italy)

Content

Context, Objectives

The COST action INTEGRAPE (CA 17111), launched in September 2018, aims at fostering open data and open science in the grapevine community. In order to set up the first set of priorities of the action, the core group has developed several actions:

- A survey sent by Email to the MC members of the project to map landscape of their local or national bioinformatics and computing resources and their practices in terms of data management
- A “World Café” was organized during the first general assembly and international conference to establish the needs of the participants and their aspirations/expectations of the action and its results. (MC + international experts)

Methods

The survey was addressed to the European partners only while the World Café included international attendees and experts.

Survey

Below the questions that were sent to the MC members of the COST action in January 2019.

Part A. Getting the broad picture

To get an initial idea regarding the resources and needs of the COST Action network in terms of support toward data management, data integration and data computing, we would like to ask each

1. To your knowledge, does your country or Institute have a policy in term of Open data and data management? (if yes: please give a reference)
2. To your knowledge, does your country or Institute have a policy in term of access to computing infrastructure? (if yes: please give a reference)
3. To your knowledge, is your country or Institute involved in large national or international initiatives in bioinformatics or open data (e.g. ELIXIR, LifeWatch, …)? If yes: is it useful for your needs?
4. What database do you use for your own data and who is in charge of its maintenance?
5. What computing infrastructure do you use for your own computations and who is in charge of its maintenance?
6. What would be the priority in terms of training for your institute?
7. What would be the priority in terms of access to bioinformatics infrastructure for your institute?

Part B. Getting more information on data produced and data management practices

1. What types/volumes of genomics-related data are you currently producing?
2. How do you expect this to change over the next 5 years?
3. How do you currently publish the data? Do you publish them at all?
4. What are the obstacles to publishing the data?
5. Are there data types you cannot publish?
6. When you seek to use relevant published data, is it FAIR (Findable, Accessible, Interoperable, Re-usable)? If not, why?
7. Are there existing norms (use of repositories, information standards, publishing conventions etc.) in the area of data publishing that you adhere to?
8. Are there candidate norms that you would like the community to adopt?
9. Are there new norms that the community could usefully develop in the framework of a COST action?
10. 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?

World Café

Eight groups of discussion were set up each with the same set of questions for two rounds of 20-30min of discussions. Paper boards were provided and pictures of the paper boards were taken.

Two first rounds of discussions

People were allowed to move from between groups to maximize their contribution to discussions. Eight reporters were kept static: their role was to capture the group discussion and to report (no slides; 5 minutes each) in a plenary session after the two rounds.

Round 1: Data:

- What types of data are you and your collaborators generating? How will they be stored (.csv, .xls, images, …)?
- What new types of data do you think will impact upon your work in future? How important is/will be automated phenotyping in grapevine research?
- What data types are specific to grape, which might be absent for other plant/crop/species? How do these data connect to other, more generic data types?
- How important is the access of data not linked to a peer reviewed publication, which otherwise never would be published?

Round 2: Integration

- What types of data do you need to combine in your work, and why? i.e. what are the common use cases for data integration?
- What are the barriers to data integration?
- What are further identifiers other than the observed object name to retrace data recording and create reliable data linking?
- What are the technical and social barriers to data sharing/data publication?
- What information systems have features you like, and why?

Third and fourth rounds of discussions

The participants were assigned to the eight groups and were not allowed to move from one group to another. A reporter was designated in each group. Again, the outputs of each groups were shortly presented in plenary session after the two rounds.

Round 3: Grape Information System

- What would you like a Grape Information System to do?
- What are your top five ranked priorities for implementation?

Round 4: Dissemination

- Why are we spreading certain information?
- Who are the potential beneficiaries of the project results?
- Which dissemination activities are appropriate to which target group?
- What are the most appropriate channels of project dissemination?
- When should certain dissemination activities take place?

Results

Results of the survey

The raw answers of the partners are presented in annexes 1 and 2.

Part A. Getting the broad picture

Partners from 17 European countries and 22 institutes answered to this part of the survey (Annex 1). In most of the institute, partners are aware of the existence of a policy for open data either at the country level and/or at the Institute level and even more have access to computing resources (Figure 1)

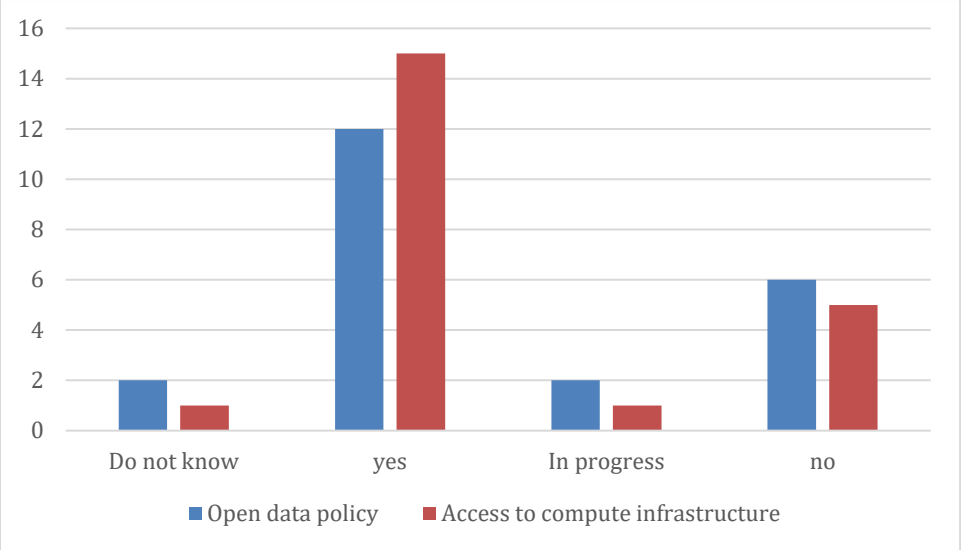


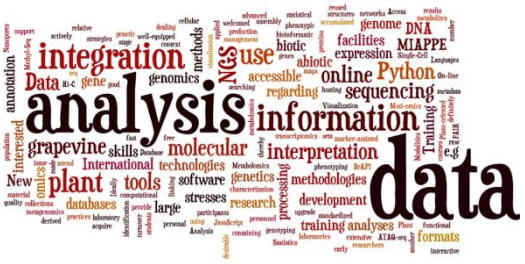
Figure 1. Answers of the partners on the existence of an Open Data Policy in their institute and on their access to an infrastructure for computing.

Many partners mentioned that their institute or country is involved in European e-infrastructures or has strong collaborations with them: 11 institutes/countries involved in ELIXIR (https://elixir-europe.org/), 2 institutes in LifeWatch (https://www.lifewatch.eu/) +/- GBIF (https://www.gbif.org/), 1 institute in DISSCO (https://dissco.eu/), 1 institutes in PRACE (http://www.prace-ri.eu/), 1 institute with EMBL (https://www.embl.org/) to cite the most important ones regarding the Integrape objectives.

In their daily work, the partners mainly use databases maintained internally or sometimes with the help of an institutional central IT service to store their data. One partner uses tools provided by the FAIRDom Hub, wich is a EU resource suporting FAIR data management. Three partners only did not use any database. For data publication in common archives or information systems, the most cited information systems were the Vitis International Variety Catalog (VIVC) and the European Vitis Database (Vitis-eu) followed by the INSDC archives.

Daily computations rely on internal servers for (9 partners), on the access to high performance clusters provided at the institutional or country level (5 partners), on the access to national super computing infrastructures (2 partners) and to a commercial cloud (1 partner). Eleven partners mentioned no computational needs.

The priority of the partners in terms of training is clearly on genomic data analysis (NGS for different purposes mainly) and to a lesser extent data standardization. The priorities in term of trainees is mostly students but in several cases management was also mentioned. Financial support to allow students to attend courses out of their institute is mentioned several times as an issue and online training material is often mentioned as a priority.



Finally, the most frequently mentioned priority in terms of bioinformatics infrastructure is the access to reliable and stable

tools for omics data analysis. National and transnational access to computing resources or to storage capacity, help for the development of an open data policy and RGPD policy compliant infrastructure were also mentioned once.

Part B. Getting more information on data produced and data management practices
Twelve institutes out of 12 countries answered to this part of the survey.
Without surprise, the partners produce very diverse types of data and in diverse volumes:

- Metabolomics data on 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

In the next 5 years, the evolutions they foresee are mainly an increase of the pace and volume of data generation but also new opportunities for science linked to technical changes (e.g. pangenomics, single cell RNAseq, RiboSeq,…). These new types of data are raising questions about references and standards for their management and analysis.

Five out of 12 partners see no technical nor sociological obstacles to publishing their data. Most of the others mention the peer review process for the publication of papers associated with the data as a sociological bottleneck. Some mentioned the lack of proper standards associated with some data types or technical issues for data publication in international archives and the lack of dedicated and skilled persons to complete these tasks. Finally, the Nagoya protocol was mentioned by one partner (legal issue).
The bottlenecks to reusing published data are a lack of proper identification of key conceptual entities (e.g. wrong or evolving geneIDs, unclear or obsolete identification of the plant material), metadata that is either incomplete or in a challenging format for re-use, provenance data that is incomplete (e.g. primer sequence) and low data quality. One partner mentioned the fact that the data is now always published along with the papers. The metadata used by the partners are mainly provided by the archives to which the data has been submitted). The questions about norms that would be useful to enforce or missing were poorly answered by the partners and difficult to report properly in terms of priority.
Finally, all partner would welcome a common hub for grapevine data and would be keen to contribute to it.

Outputs of the World Café

Data

Genotpying (PCR, SSRs, SNPs)
Multi-omics (epigenomics, transcriptomics (developmental stages), proteomics)
QTLs and allele specific data
Micro RNA data
Tageted metabolomics.

Metagenomics is growing in importance
Chomatography, spectral data at various scales (GCMS, LCMS, NMR, UV-Vis)
Time series

What's unqique about grape? Diversity of data used, and its application in food/wine.
Rootstocks, clones.
Biotic stresses
Soil, climate, microclimate data (abiotic stresses - drought, temparature, CO2)
GPS, slope, agronomical data (training, density, etc..)

Data publication should be encouraged, but unreviewed data may be rubbish; data curation is important
Phenotyping data: a lot of it is still not high throughput. However there is a growing wealth of spectral data.

Integration

Want to integrate different omics data sets
Integration of omics, non-omics data, phenotypes, data about the environment and data about the management of the experiment

Need to track sample, DNA, tissue, developmental stage, data sets - need machine and human readable formats, and appropriate identifiers (e.g. various chemical identifiers).

Need to develop recommendation of experimental design in a modular way (depending on the type of question to address)

Need for bioinformatic and statistical methods (or training on existing methods) : to integrate heterogeneous data or even to integrate same types of data acquired on different platforms.

Barriers:
Complexity of methods, protocols, statistical parameters.
Incompatibility of different technologies.
Software, poor UI
Absence of standardized formats, ontologies, published workflows and an integration system accessibility of data (not even a list of available data)
Poor fit of existing tools to breeding use case
Lack of communication; competition; absence of a reward system for integration; commercial sensitivity
Lack of training
Should support pre-release data sharing among collaborators.
Support of Bermuda principles for early data release.
Inconsistency of identifiers across disciplins/data silos
Dispersion of the data: difficult to find all the relevant and existing public data
Lack of standardization of the vocabularies
Broken links between metadata and data, poor metadata (e.g. no unit indication!)
Lack of description of the platforms/sensors/tools used to produce the data or of the methods used to normalize the data when normalized

Good models include CRiBi, InterPro, NCBI, Panther, TAIR

GrapeIS

Things people wanted (rankings in brackets) - I have only merged where the overlap was very clear

Web pages for the grapevine community regularly updated with relevant news and links (1)
Meta-analysis (1)
User-friendliness (1)
List of key resources: reference data sets (e.g. genomes), repositories, standards, .. (1, 2, 4)
Reference genes (1, 2) and germplasm (1)
Links to external repositories (e.g. NCBI (2)
Allow data/metadata upload (2, 3)
Glossary of terms across disciplins (3)
Ontologies (2, 3) - new and recommended existing
Community standards (5)
Consistent annotation (3)
Workflows and tools (3)
Specification of data formats (4)
Centred around key object types (gene, genotype, environment, developmental stage, tissue) (4)
Checklist for standardized metadata (4)
Validation (4)
Integrated search, search tool (4, 5, 5) - depends on metadata, self-curated or centrally curated?
Visualisation tools (5)
Atlas (5)
Datatype list (6)

Dissemination

Aims: favour open science and collaborations
Avoid competition
Facilitate outreach
Facilitate future funding

Beneficiaries: plant scientists, early career stage researchers, teachers, public, regulators, equipment manufacturers, wine industry (farmers/wineries), consumers
Means: conferences, courses, social media, professional materials (e.g. animations), e-learning, wine tasting, open days, training, workshops, dedicated pages on a web site
Guidelines
Papers in scientific journals
Papers in specialized magazines

Metrics on dissemination ?