Evolving the VO: from interoperable data collections to an integrated system of services for data-intensive science

Fabio Pasian[*], Marco Molinaro[†], and Giuliano Taffoni[‡]

INAF – Osservatorio Astronomico di Trieste, Trieste, Italy

Abstract

The Virtual Observatory (VO) represents a successful international enterprise providing interoperability of data collections, thus allowing the possibility of multi-frequency and multi-messenger research. The Big Data era, that astrophysics has stepped into, is forcing scientists to perform data-intensive research. This new concept requires an evolution of the VO concept to provide additional services, in order to transform the VO: from sets of interoperable data to an integrated system of services capable of supporting data-intensive science.

Keywords: *Virtual Observatory, big data, data-intensive research, science platforms, open science clouds*

1. Introduction: the Virtual Observatory

1.1. The VO

The Virtual Observatory (VO) is the vision that astronomical datasets and other resources should work as a seamless whole. The original ambition of the VO was linking all of the main activities of astronomers into a coherent "circular" framework: from publications (with their associated data, tables and figures) to new observing proposals, from observatory info (meteo, calibrations, raw data) to analysis (data processing software, catalogues, data products) from comparison with theory and models, back to publications. Although extremely successful, the VO achieved results only partially fitting the above-described original ambition.

The VO aims at maximising scientific results out of archival research, by providing data of the highest quality achievable to all scientists and interested individuals; and also achieving a new type of science, crossing not only the boundaries of nations, but also those of wavelength, messenger, instrument specificities. In this way, science is not enabled just for power users, but the full research community is expected to have meaningful access to all data.

The mechanism governing the VO is most easily described by analogy with the World Wide Web: the information (and specifically data, in the VO case) is expected to appear seamlessly at the user's desk. And just as in the case of the WWW, the VO is not a fixed system, but rather a way of doing things.

The VO is made possible by the standardisation of data and metadata, by the standardisation of data exchange methods, and by the use of a registry, a repository of resources (entities, standards, data collections, services, ...), their mutual relationships and actionable interfaces.

The VO was probably the first environment to anticipate in practice the eventual implementation of the so-called FAIR principles Wilkinson et al. (2016): data and services provided are Findable, Accessible, Interoperable and Re-usable.

[*] fabio.pasian@inaf.it, IVOA Exec member, former IVOA Chair, Corresponding author

[†] marco.molinaro.inaf.it, IVOA Data Access Layer (DAL) WG Chair

[‡] giuliano.taffoni@inaf.it, IVOA Grid and Web Services (GWS) WG Chair

DOI: 10.52526/25792776-2020.67.2-302

1.2. The IVOA

Since 2002, many projects and data centres worldwide have been working towards the goal of implementing an increasingly effective VO. The International Virtual Observatory Alliance (IVOA) is the organisation that debates and agrees the technical standards that are needed to make the VO possible. It also acts as a focus for VO aspirations, a framework for discussing and sharing VO ideas and technology, and body for promoting and publicising the VO. Participation in the IVOA includes VO initiatives in 19 countries (plus The Netherlands and Thailand interested in joining), 1 super-national institution (the European Space Agency), and 1 international collaboration (the European VO initiative).

The aims of the IVOA standards are:

- interrogate multiple data centres in a seamless and transparent way;
- new powerful analysis and visualisation tools within that system;
- a standard framework for data centres to publish and deliver services using their data.

Lots of successful work has been done. But it is to noted that IVOA standards have always been driven by data resource interoperability. A limited amount of work has been therefore done on processing/analysis services (and therefore a limited number of services is available in that area).

For an overview of VO and IVOA activities, see e.g. the collection of papers edited some years ago by Hanisch (2015).

2. The VO and Big Data

With the increasing size of data sets in the "Big Data" era, to move data around the network becomes heavy and cumbersome.

To explain and visualise the problem, let's consider as an example the classification of star spectra in globular clusters. The user (see Figure 1) finds two VO-compliant archives containing digitised objective prism plates and spectra, respectively. The archives do not provide what is needed (i.e. catalogues of classified stars), but data retrieval services only. The user thus needs to download images and spectra (with delays caused by the network) and perform the processing at his/her own premises.

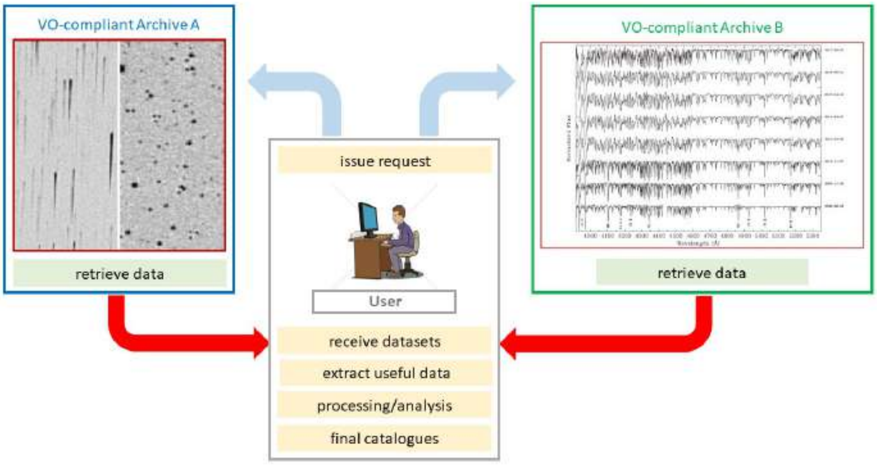

Figure 1. Classification of star spectra: Data Centres provide retrieval services only (explanation in the text). Yellow boxes indicate user actions and code, green boxes server-side actions.

How can this situation be improved? The provision of increased services (involving some level of computing) at the Data Centres would help a lot.

The need for a tighter connection between data and computing facilities has been brought forward for more than a decade. This was initially referred to the connection between the VO and Grid computing by Pasian et al. (2007), and to data mining in the VO context by Pasian et al. (2012). This new paradigm has gained momentum in the IVOA in the past 5 years as *”bring processing to the data”*. This topic was included in 2016 by the IVOA Science Priorities Committee (after discussion) among the top priorities for the VO. But the IVOA is a ”joint venture” of separate national VO initiatives with no budget of its own, and since each initiative has its own set of priorities and funded activities, progress in this direction has been limited.

But recently, the importance of providing data processing and analysis services as part of archival activities has gained momentum. It is therefore foreseen that ”bringing processing near the data” will be increasingly common. Of course, computing services applied to archived data can occur in a number of different ways.

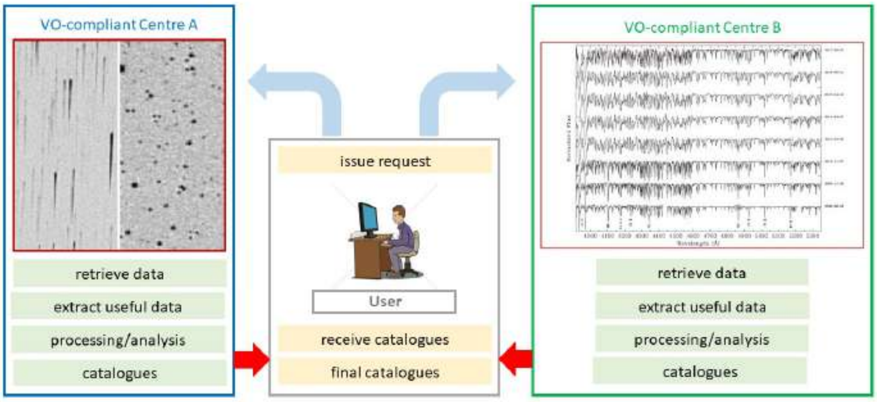

Figure 2. Classification of star spectra: Data Centres provide complete computing and classification services (explanation in the text). Yellow boxes indicate user actions and code, green boxes server-side actions.

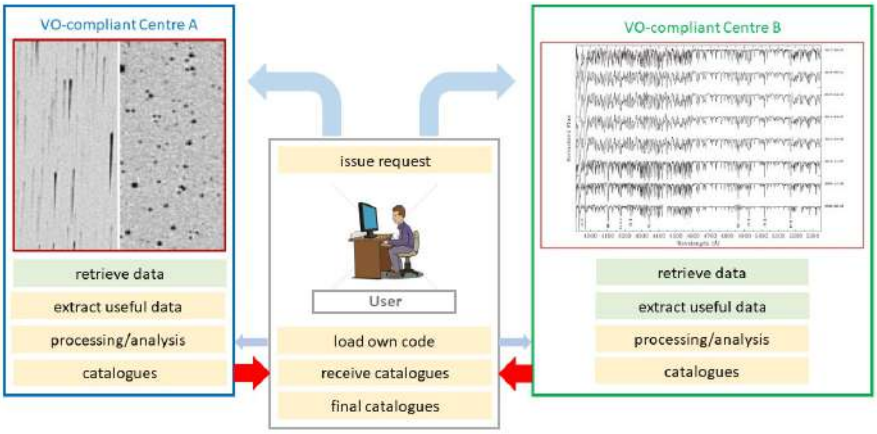

Figure 3. Classification of star spectra: Data Centres provide the running of user-provided code as a service (explanation in the text). Yellow boxes indicate user actions and code, green boxes server-side actions.

Proceeding with the example proposed above (classification of star spectra), new sets of services can be provided by Data Centres (Figure 2). E.g. cut-out of spectra in spectral images (e.g. using SODA, the IVOA standard Server-side Operations for Data Access web service capability that can act upon the data files, performing various kinds of operations: filtering/subsection, transformations, pixel operations, and applying functions to the data); and, in increasing order of computing complexity: optimal extraction of spectra, classification of spectra, preparation of catalogues of classified spectra. Only final results (catalogues) need to be sent back through the network. This case, in which users can run available services at the remote site can be described in Cloud terms as Software-as-a-Service (SaaS).

But Data Centres could also provide as a service the running of user-provided code (Figure 3). E.g. for the extraction of spectra or classification of spectra (e.g. the user could try new algorithms for this purpose). Also in this case only final results (catalogues) need to be sent back through the network; but this solution allows processing to be performed directly by users (and not on their behalf). In Cloud terms, this mechanism can be described as Infrastructure-as-a-Service or Platform-as-a-Service (IaaS/PaaS). While maybe better for the remote users, this mechanism implies a more complex handling of Authorisation&Accounting for a Data Centre. The increased complexity of the Autentication&Authorization&Accounting affects only marginally the technical implementations, but rather the policy for the usage of computing resources offered by an institution to external users for their own research. This implies a revolution in the approach of Data Centres and institutions on the way resources are consumed and new ways to report to funding agencies: a global agreement could be found in the framework of an Open Science Cloud (as discussed in the next sections).

3. Extending the VO paradigm

3.1. Concepts

The requirements of the Astronomy community (especially access to archives and the VO, processing of data, "bring computing to the data") call for an expansion of the standard concept of the VO. A tighter integration (or at least interoperability) of data and computing resources is needed.

The key point is that Data Centres, ideally linked within an International Open Science Cloud, are to provide to all of their users (following a VO-aware approach) the following resources:

1) archived dataset collections;

2) an amount of computing power;

3) user storage.

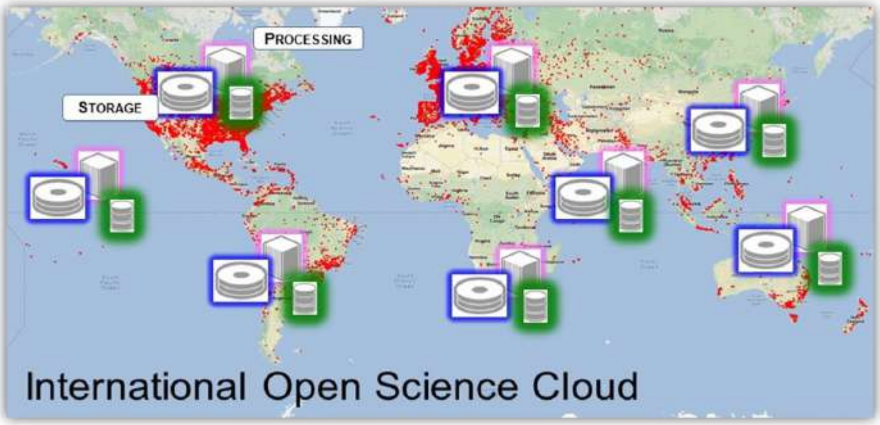

Figure 4. The concept of International Open Science Cloud. Each Data Centre provides the archived dataset collections, an amount of computing power, and user storage: all the Data Centres are made interoperable through federation of Open Science Clouds (viewgraph shown by David Schade, CADC, at the UN/Italy Open Universe Workshop, Vienna, 22 Nov 2017).

Building an International Open Science Cloud (Figure 4) is a vision shared by many colleagues all over the world, e.g. the Astrocloud initiative set up by VO-China, described by Cui et al. (2017). It is furthermore to be noted that the EGI (and now EOSC) in Europe and CANFAR Compute-Canada clouds have a similar approach and have worked on interoperability / federation of clouds (through a collaboration between INAF-OATs and CADC), as discussed in Bertocco et al. (2018) and Bertocco et al. (2020).

An International Open Science Cloud scenario allows to encompass different needs (Figure 5):

- Large collaborations project-oriented data centres offer access to instrument specific data (e.g. SKA, LSST). They build their own data and computing on top of dedicated infrastructure or Private or Hybrid Clouds (IaaS).
- Multi-purpose Data Centres: to manage small archival needs: surveys, key projects, thematic data resource aggregations, multiple projects, etc.
- Individual scientists.
- Citizen Science and Education / Public Outreach.

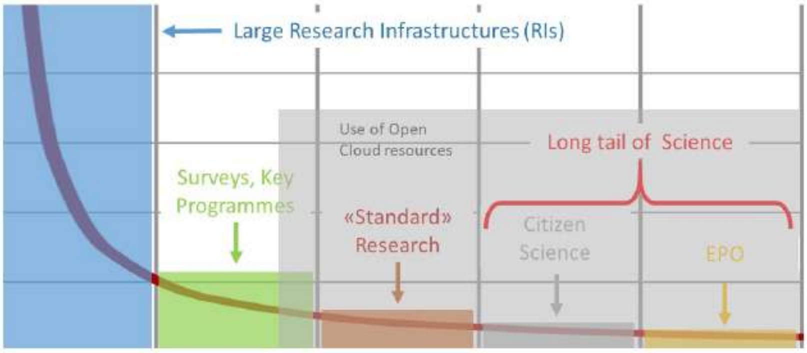

Figure 5. Usage of computing and data resources by different classes of users: in grey, the area where Open Science Clouds prove to be most useful.

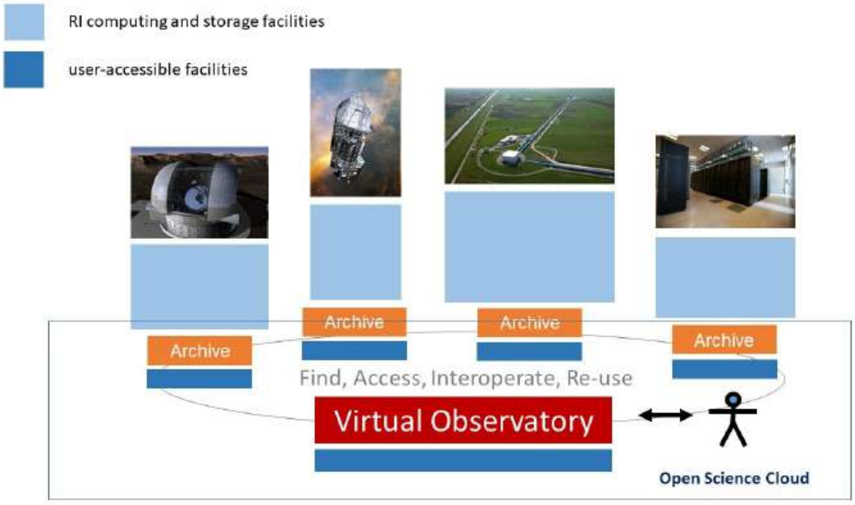
RI computing and storage facilities
user-accessible facilities
Archive
Archive
Archive
Archive
Find, Access, Interoperate, Re-use
Virtual Observatory
Open Science Cloud

Figure 6. Open Science Cloud from the users' point of view (explanation in the text).

Figure 6 depicts the situation from the users' point of view. The observational Research Infrastructures (ground-based, space-borne, multi-messenger, including models and theory data providers) have their own infrastructure and store their data in their own archives. Additionally, personal space and some computing resource is provided to users, who can combine their own resources with what is available within an Open Science Cloud. The Virtual Observatory needs to be embedded in such a cloud, since it is the key to provide FAIR access (through community-accepted standards) to all available data and computing services.

3.2. Technical feasibility

Most of the technical tools needed for this switch in paradigm are already available. Data centres should offer computing and storage capabilities to process data. Where? Computing resources must be close to data: the advent of 100 Gbps networks (a speed comparable with the access speed of magnetic disks) promotes the concept of network proximity (the concept of ”data lakes” with ”computing islands”). Data providers may offer the possibility to process data using Web Apps within a SaaS framework. Additionally, astronomers should be able to deploy their own code in a PaaS mode: the Science Platforms approach.

A Science Platform (SP) is an environment designed to offer users a smoother experience when interacting with data and computing resources. A SP implements SaaS and PaaS capabilities to allow the most ”elastic” use of the resources and of the data. A SP provides services to search data and software, process data with software (reduction, analysis, visualisation, etc.), access to computing resources, access to storage resources.

This point is worth expanding. SPs actually fill a gap in the services provided by Open Science Clouds, that currently offer only IaaS, while public/commercial Clouds often offer PaaS as well: in such environments, building SPs would probably be simplified by the additional services provided.

Deeply embedded with the SP concept is the mechanism of containers, which allow the practical implementation of moving the code to data. It involves encapsulating or packaging up software code and all its dependencies so that it can run uniformly and consistently on any infrastructure. Containers are a portable, lightweight, efficient and easy to maintain solution to offer software: they are at the basis of scientific experiment reproducibility; containers, just as data, should be annotated in order to make them FAIR and associated to scientific data (such connectors allow users to check their reciprocal compatibility).

Among the current technical tools allowing the use of containers, one may mention Docker (for micro servicing, default standard, isolation, etc.) and Singularity (for advanced environments, HPC, HTC), and the so-called Orchestrators (e.g. Kubernates).

Popular terms connected to container technology are the following.

- *Registry* “hosts” containers and allows to easily maintain and deploy on infrastructures (this Registry refers to containers, and is of course *different* from the VO Registry).
- *Marketplace* provides the tools necessary for the communities to share their science products in a harmonised way respecting the FAIR principles.
- *Annotations* (metadata), *versioning* and *DOI* allow precise citation of the used software on dataset. The use of keywords allows to filter and organise the content.

3.3. Steps ahead and challenges

An important step ahead could be the definition of AaaS as a new framework for analysis. Analytics-as-a-Service (AaaS) provides subscription-based data analytics software and procedures through the cloud. AaaS uses data mining, predictive analytics and AI to effectively reveal trends and insights from existing data sets. Data Centres (maybe large project data centres) may look to AaaS as a specific service to explore and analyse their data offering it as a WEB applications (please note that ”WEB” does not mean ”graphical browser”). AaaS implies access to high end computing resources (HPC, GPUs etc). For details, look out for Taffoni (2021), in preparation.

The main challenges to be tackled are listed in the following.

- The identification of software tools and packages for data processing and/or the deployment of custom workflows to the platform.
- How to take advantage of HPC and HTC computing infrastructures that require batch processing to execute analysis.

- Tools, software, computing and storage resources should be findable, accessible and sharable transparently.
- Authentication, Authorisation and Accounting: Data Centres want to know who is using their infrastructure and why!
- The identification of API and protocols to build intelligent data lakes capable of distributing data among different cloud Data Centres and offering services for SPs in order to couple software, computing, and data.

4. The role of IVOA

The IVOA, as the organisation defining the VO standards, has the key role to play in the evolution of the VO and the extension of its paradigm. The main items to tackle are the following.

1) The current IVOA standards provide protocols for data discovery and access.

2) The IVOA shall identify metadata to describe software: a software "data-model".

3) The IVOA shall identify technologies to build software marketplaces based on "common" standards.

4) The IVOA should discuss, build and promote policies and trust to implement a "data lake with computing islands" (providing services consuming resources requires important policy decisions by each data centre).

5) The IVOA shall investigate a way to let public data to remain fully accessible, "unharmed" by Authentication&Authorisation policy concerns. Up to now, VO-compliant data centres have provided free and public access to data (usually by means of anonymous access). Additional services may imply a set of different requirements, therefore it must be possible to enforce A&A&A (Authentication-Authorisation-Accounting) e.g. for complex services and user processing, while keeping at the same time anonymous access, e.g. for read-only access and no processing, EPO, citizen science, simple services requiring computing.

A promising step made by IVOA to cope with these needs is the fact that the IVOA Technical Coordination Group (TCG) has included the update to A&A&A technologies, the software "data-model" and a discussion on Science Platforms in the 2020 IVOA Technical Roadmap.

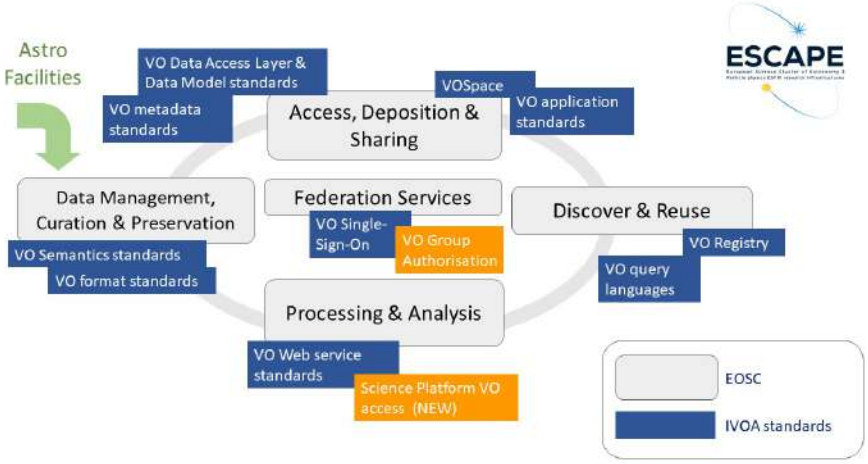

Figure 7. Integration of astronomy VO data and services into the European Open Science Cloud (EOSC).

As an example of practical work being carried out, the ESCAPE project (funded by the European Union within its Horizon 2020 programme) can be mentioned. The developments relevant to this discussion are carried out within its VO (Virtual Observatory) and ESAP (ESFRI Science Analysis Platforms) work-packages, and are shown in Figure 7. In the figure, the tools and mechanisms provided by EOSC cloud to perform a set of operations are marked in grey. The blue boxes represent existing VO standards, which need to be interfaced with the EOSC tools; the orange boxes define the need for new VO standards, that need to be defined.

The definition of a software "data-model" (item 2 above) could be performed in connection with the already advanced work provided by ASCL. It is worth noting that ESCAPE (within its OSSR, Open Software and Services Repository work-package) is also tackling code description with respect to the data formats it can consume.

The above example shows that new IVOA standards are needed to support the extensions.

- Besides VO Authentication (handled through the IVOA Single Sign-On standard), to use resources Authorisation is needed as well (at the level of group authorisation). This is being discussed within IVOA GWS WG since October 2016, and the approval of an IVOA standard is expected soon.
- VO-compliant access to Science Platforms. This is a new concept, being worked upon within ESCAPE project; a new set of standards will be proposed to the IVOA.

5. Conclusions

Recently the importance of providing data processing and analysis services as part of archival activities, and the concept of an International Open Science Cloud, have received increasing interest.

- In early 2019, among the 294 "APC" (Activities, Projects, and State of the Profession Considerations) white papers submitted by astronomers to the Astro2020 Decadal Survey on Astronomy and Astrophysics organized by the US National Academy of Sciences, one by Desai et al. (2020) is particularly relevant to this discussion. The white paper advocates for the adequate funding of data centers to develop and operate "science platforms", which will provide storage and computing resources for the astronomical community to run analyses near the data; these platforms should be furthermore connected among each other to enable cross-center analysis and processing.
- The idea of building a multi-disciplinary Global Open Science Cloud (GOSC) was initiated during the CODATA 2019 Beijing conference. A Global Open Science Cloud Workshop has been organised for 3-4 November 2020 with the participation of representatives of international initiatives, research communities and public digital infrastructure providers. The IVOA will be represented by its current Chair.

In this framework, it is essential for the Virtual Observatory to evolve its concept in order to allow data-intensive research in the Big Data era.

From the technical point of view, there no obstacles: the Cloud paradigm offers a good set of solutions. The main issue of allowing the various clouds to operate together (i.e. federation) has been successfully experimented in a number of cases.

However, for this new evolution of the VO to be accepted and become operational, there are a number of policy implications, and decisions to be taken:

- all Data Centres are to provide, besides access to archive data, personal user storage and computing resources (here there are also some local technical decisions to be made);
- the VO community needs to adapt and evolve the IVOA standards to support this new paradigm: an agreement on the new technical standards to be defined needs to be found;
- the international astronomical community needs to move coherently in this direction: and here the lead of governing bodies to push these policy decisions forward is necessary.

Acknowledgements

The authors wish to thank the colleagues who shared their visions and points of view on Open Science Clouds (in particular David Schade and Chenzhou Cui), and the whole team at INAF-OATs working on VO and computing. The European Science Cluster of Astronomy Particle Physics ESFRI Research Infrastructures (ESCAPE) project is funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement no. 824064.

References

Bertocco S., Dowler P., Gaudet S., et al. 2018, Astronomy and Computing, 24, 36

Bertocco S., Major B., Dowler P., et al. 2020, in Astronomical Data Analysis Software and Systems XXVII. Astronomical Society of the Pacific, p. 327

Cui C., Yu C., Xiao J., et al. 2017, in Astronomical Data Analysis Software and Systems XXV. Astronomical Society of the Pacific, p. 553

Desai V., Allen M., Arviset C., et al. 2020, Bulletin of the AAS, 51

Hanisch R. J., 2015, Astronomy and Computing, 11, 73

Pasian F., Taffoni G., Vuerli C., 2007, Highlights of Astronomy, 14, 601

Pasian F., Brescia M., Longo G., 2012, in Science - Image in Action. World Scientific Publishing Co. Pte. Ltd., p. 230

Taffoni G., 2021, in Astronomical Data Analysis Software and Systems XXX. Astronomical Society of the Pacific, pp invited, in preparation

Wilkinson M. D., Dumontier M., Aalbersgberg I. J., et al. 2016, Scientific Data, 3