SF2A 2021
A. Siebert, K. Baillié, E. Lagadec, N. Lagarde, J. Malzac, J.-B. Marquette, M. N'Diaye, J. Richard, O. Venot (eds)

OPEN ACCESS TO SCIENTIFIC DATA: EXCERPTS FROM THE INSU PROSPECTIVE

C. Bot[1], F. André[2], M. Allen[1], F. Bonnarel[1], A. Chambodut[3], S. Galle[4], F. Genova[1], M. Gérin-Laslier[5], F. Huynh[6], H. Pedersen[7], V. Stoll[8] and J. Vergne[3]

Abstract. Data is at the heart of the scientific method at INSU, the French national institute for Universe Sciences. INSU disciplines were engaged in data sharing and data management long before political acceptance to Open Science and the definition of FAIR principles (Findable, Accessible, Interoperable, Reusable). This early involvement brought us to the leading edge of data sharing. Open access to scientific data was therefore a natural topic for the first INSU inter-displinary prospective organized in 2019-2020. The discussion was organized on different points: the FAIR context, scientific data management and services, data models and metadata standardisation, and the certification of data repositories (CoreTrustSeal). The current paper provides an excerpt from the discussions, conclusions and recommendations, with an intended bias toward astronomy and astrophysics. The aim is to trigger interest and give extra motivation to read the full online document with the conclusions from the prospective on open access to scientific data.

Keywords: open science, data sharing, FAIR, data models, metadata, certification

1 Introduction

INSU (Institut National des Sciences de l'Univers) is the French national institute of CNRS on Universe Science and includes 4 different research domains pertaining to Earth Science as well as Astronomy and Astrophysics. In 2019-2020, INSU organized its first inter-displinary prospective exercise among these different disciplines.

Data is at the heart of research in all INSU disciplines and we are today at the forefront of data sharing, thanks to a long history of involvement, long before Open Science was a hot topic. Therefore, open access to scientific data was easily identified as one of the inter-disciplinary challenges to be discussed as part of the INSU prospective. A workshop on these questions was organized in Strasbourg in January 2020. A recording of the discussions (in French) is available online*. The conclusions and recommendations that stemmed from the prospective are gathered in an online document†. The current paper aims at summarizing some of these discussions and recommendations, with a bias on those pertaining to astronomy and astrophysics. The intent is to trigger interest for researchers in these fields and give some additional motivation to read the full document.

Let's step back. Why are researchers interested and willing to invest time on Open Science and share data? Open Science for scientific data is a "hot topic", with numerous political demands. One pragmatic incentive to share data is that it is now becoming mandatory in some contexts (e.g. for data obtained as part of H2020, ERC or ANR projects). Another motivation for data sharing is more general: it gives trust, shows reliability and accountability. Last but not least, sharing data opens new research fields.

[1] Observatoire Astronomique de Strasbourg, Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg (ObAS), UMR 7550, 67000 Strasbourg, France

[2] Observatoire Midi-Pyrenees, Toulouse, France

[3] Université de Strasbourg, CNRS, Institut Terre et Environnement de Strasbourg, ITES UMR 7063, Strasbourg F-67084, France

[4] Université Grenoble Alpes, CNRS, IRD, Grenoble-INP, IGE, 38000 Grenoble, France

[5] LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Universités, 75014, Paris, France

[6] Institut de Recherche pour le Développement, Marseille, France

[7] Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, IRD, IFSTTAR, ISTerre, 38000 Grenoble, France

[8] Observatoire de Paris, PSL Research University, Paris, France

*http://www.canalc2.tv/video/15656

†https://extra.core-cloud.net/collaborations/ProspectiveTransverseINSU2020/Defi-14/Documents%20partages/defi14_final.pdf

2 FAIR context

Key concepts of data sharing and open science have been condensed in the acronym FAIR (Wilkinson et al. 2016), which stands for Findable, Accessible, Interoperable and Reusable. These FAIR principles describe criteria that should be met for data to be open. The FAIR concept is evolving in a larger context with a lot of developments at different levels happening now. The European commission is providing a strong support to EOSC (European Open Science Cloud European Open Science Cloud 2018). At the international level, the Research Data Alliance (RDA) provides a neutral forum to discuss all aspects of data sharing and there is a French RDA node. At the national level, CNRS published a roadmap for open science (Centre National pour la Recherche Scientifique 2019) and the ministry for research and education have a national plan for open science (Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation 2018). The prospective exercise highlighted that it is important to participate in activities like EOSC to make sure that what comes out of these initiatives will meet the needs of our communities and that our major services will be included.

At the heart of FAIR principles, the idea is to do science more efficiently. Yet sharing data requires a significant work at the level of each discipline. In practice, this means being on the same page as far as formats, metadata or exchange protocols are concerned. Astronomy was a pioneer on these aspects with the Virtual Observatory standards. An example of such standards, the provenance data model, is being presented by Servillat (2020) in this same volume. Today, data sharing is in the mindset of all disciplines. FAIR is a revolution but one has to keep in mind that it comes at a cost. One of the out-coming messages of the prospective is that not all data is intended to be FAIR. The consequence is that priorities have to be set for which data should be FAIR. Data from research infrastructures (European Strategy Forum on Research Infrastructures 2018) should be FAIR, as well as the ones from National Services (SNO- Services Nationaux d'Observation) or data associated to publications. For other data sets, there is a need to put priorities and think about whether these data should be FAIR or not and for how long. Inter-disciplinary data sharing also has a cost and a recommendation from the prospective was to start with use cases and existing frameworks (e.g. gravitational waves).

3 Management of scientific data and services

The discussion on scientific data management was done by considering two distinct yet complementary categories: data managed by data sharing platforms, observatories or infrastructures on the one hand, and any other datasets on the other hand (long tail, data attached to publications, data outside SNOs, non-digital data).

Data management is a key aspect of sharing data and enabling open science. In order for scientific data to be used and used again, one has to assign metadata and use inter-operability standards and frameworks. In this work, scientists, software engineers and documentalists have complementary roles. In France and for INSU, data management is done through two research infrastructures working (CDS -Centre de Données de Strasbourg- for astronomy and DataTerra for earth system) and the structure of OSUs (Observatoires des Sciences de l'Univers) and national services (SNOs). Recommendations made by the prospective are that this national structure within INSU is an asset and it should be used to suggest EOSC services that suit our usage. Another recommendation is that there is a need for a support of infrastructures, to favor thematic repositories. Finally, to avoid the risk of duplication of data at different places, repositories should be federated by harvesting metadata.

For data outside observatory services (long tail, non-digital data, ...), the issue is the deterioration of data with time, even if data is associated to a publication (Pepe et al. 2014). Data management is the responsibility of data producers. One of the difficulties to take care of these data is to have sustainable funding (for times longer than the time of the project). One proposition is to have overheads on projects (around 10%), especially for projects which would require a lot of data management on timescales larger than the project duration. Another proposition from the prospective is to put priorities on long tail data sets but that the description should be done beforehand for all data sets. We also recommended to do an inventory of data services in order to guide researchers and document the need for a long tail repository. Lastly, scientists should continue to develop their culture on data management (e.g. how important it is to attach metadata on the provenance to data) and there will be data correspondents or referents within each observatory (OSU).

4 Complimentary professions and skills

Among the different questions we asked ourselves during the prospective on open access to scientific data, professions and skills were a recurring discussion item. Data sharing involves different people with different

skills (researchers, engineers, documentalists, information science specialists, legal experts, ...). New professions are arising (data analyst, data architect, data steward, data scientist, ...) and the definition of these jobs are not always clear or unequivocal. A recommendation is hence to do an analysis of the different professions needed to take care of data. There is a need to identify, recognize and value professions associated to data management and data preservation. A recommendation was made to recognize skills related to data. Working on data is a specific work that is complementary to works on algorithms or on infrastructures and IT systems.

5 Data description, formats

Standards for formats and metadata are central to interoperability. Standardisation in astronomy and astrophysics came very early compared to other disciplines. A good example is the FITS standard, which enabled an easy exchange and usage of astronomical images, spectra, ... This approach widened and formalized through the concept of Virtual Observatory and the creation of the International Virtual Observatory Alliance (IVOA), where standards for the virtual observatory are defined. Today, 45 standards exists for formats, protocols, and data models, covering different aspects of FAIR principles for different types of data. The IVOA framework is seen as the way to do interoperability and to open data in practice. In France, the Action Specifique Observatoire Virtuel (ASOV[‡]) is coordinating the French involvement and sharing good practices. Community training on these topics is important at all levels.

6 Certification of data repositories and services

All actors want to be assured that the structure giving access to their data is permanent, robust and has appropriate data management practices. Certification is the way to prove openly the trustworthiness. To do so, data repositories and services follow a set of principles gathered in the acronym TRUST: Transparency, Responsability, User Community, Sustainability, Technology (Lin et al. 2020). This certification applies to the repository, not the data quality nor its value to the community. For scientific services, the best suited certification is CoreTrustSeal[§]. Currently, in France, only two services have this certification: SISMER (sea data, Ifremer) and CDS. The prospective highlighted that the certification process involves governance. Auto-evaluation is a useful process, even if the service does not submit the certification.

7 Conclusion

Open science and data sharing are key aspects that are of interest across all INSU disciplines. The prospective organized by INSU was an occasion to discuss and reflect on the FAIR context, management of scientific data and services, and data models and meta-data harmonization. One of the overarching messages was that science should be at the heart of the process and should be the engine of evolution and strategic choices. All conclusions and recommendations are available online at `https://extra.core-cloud.net/collaborations/ProspectiveTransverseINSU2020/Defi-14/SitePages/Accueil.aspx`.

References

Centre National pour la Recherche Scientifique. 2019, Feuille de Route du CNRS pour la Science Ouverte

European Open Science Cloud. 2018, Strategic Implementation Plan

European Strategy Forum on Research Infrastructures. 2018, Strategy Report on Research Infrastructures

Lin, D., Crabtree, J., & Dillo, I. e. a. 2020, Sci Data, 7, 144

Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. 2018, Plan national pour la science ouverte

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. 2014, PLoS ONE, 9, 8

Servillat, M. 2020, in Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics

Wilkinson, M., Dumontier, M., & Aalbersberg, I. e. a. 2016, Sci Data, 3, 160018

[‡] `http://www.france-ov.org`

[§] `https://www.coretrustseal.org`