FAIR DATA THROUGH A FEDERATED CLOUD INFRASTRUCTURE: EXPLORING THE SCIENCE MESH

Research in Progress

Angelo Romasanta, ESADE Business School, Barcelona, Spain, angelokenneth.romasanta@esade.edu

Jonathan Wareham, ESADE Business School, Barcelona, Spain, jonathan.wareham@esade.edu

Abstract

Despite the many promises of cloud computing in science, its full potential is yet to be unlocked due to the lack of findable, accessible, interoperable and reusable (FAIR) data. To investigate the barriers and promises of FAIR data, we explore the Science Mesh, a federated mesh infrastructure enabling interoperability across research cloud service providers. As a starting point, the project will connect 300,000+ researchers from eight providers across Europe, Australia and beyond. Through our analysis of the Science Mesh, we frame FAIR data as a collective action problem permeating across two levels: researchers downstream and cloud service providers upstream. On one hand, the scientific community would be better off if researchers made their data FAIR, yet misaligned incentives hinder this. On the other hand, cloud providers are also not incentivized to pursue interoperability with other providers, despite its benefits to users. By addressing these two dilemmas towards FAIR data, the Science Mesh promises to unlock new ways of collaborating through frictionless sync and share, remote data analysis and collaborative applications.

Keywords: FAIR data, Interoperability, Digital infrastructure, Federated cloud, Open Science

1 Introduction

Cloud computing has enabled distributed researchers to conduct increasingly ambitious and complex experiments (Hoffa *et al.*, 2008; Berriman *et al.*, 2013; Yang *et al.*, 2014). Despite the advances it has brought to research collaborations, the cloud still has not fully realized its impact due to data not following FAIR principles – that the data and the tools to analyse them are findable, accessible, interoperable and reusable (Wilkinson *et al.*, 2016). The lack of FAIR data has led to various negative consequences including wasted time, additional storage costs, extra licence costs, research retraction, double funding, missed interdisciplinary research and lost potential economic growth (European Commission, 2019). All in all, this has led to the European Union losing an estimated total of 26 billion EUR each year due to the lack of FAIR data (European Commission, 2019). Realizing its importance, various stakeholders including policymakers, research funders and universities have then called for the scientific community to embrace good data management and data stewardship practices (Madduri *et al.*, 2019; Mendez *et al.*, 2019). However, transitioning to FAIR data can be challenging due to various individual and systemic barriers. In this article, we explore the challenges and promises of FAIR data through the lens of a digital infrastructure initiative called the Science Mesh.

The FAIR data movement grew in 2016 when its principles were published in the journal Scientific Data (Wilkinson *et al.*, 2016). With the increasing volume and complexity of data generated by the scientific community, these principles were necessary to improve the findability of existing data, ease the access to data at scale and allow reuse of data (Wise *et al.*, 2019). The most promising impact of FAIR is enabling computational systems to take existing data and find new uses with minimal human intervention. Despite its ambitious promises, transitioning to FAIR data is not straightforward. Data governance has been construed as a collective action problem where individuals would be better off if they cooperated, but are not compelled to do so to maximize their short-term personal benefits (Ostrom, 1990; Benfeldt, Persson and Madsen, 2020). If researchers properly labelled and published their data to be easily reused by others, the scientific community would benefit as a whole. However, doing so may be seen as costly for researchers, diverting their time away from their primary research.

What makes achieving FAIR data extra difficult is that this social dilemma does not only permeate at the level of individual researchers but also at the level of their cloud service providers. Particularly, we refer to the fragmentation of the cloud service landscape (Ortiz, 2011; Lewis, 2013; Ranjan, 2014). To illustrate this, while for some countries, one data service may cater to a large fraction of the research data storage needs; other larger countries may have different arrangements from region to region. Some countries might not even have any information infrastructure policy in place, making it each research institute's responsibility to avail or set up their respective cloud infrastructure. While the presence of different options enables organizations to choose the best that suits their needs, this heterogeneity also has led to a fragmented landscape with multiple private and public cloud providers who are incentivized to lock their data in, out of reach from other providers.

While solutions have been taken by individual researchers and individual projects, solutions at larger scales are still lacking. To address this, the initiative investigated in this article, the Science Mesh (sciencemesh.io), intends to connect existing research data services through vendor-neutral APIs and protocols. The Science Mesh will bring together existing services that currently host 300,000+ user accounts across Europe and beyond. After this initial phase, the infrastructure will be expanded to serve the rest of the research community. It will be a key component of the European Open Science Cloud, an initiative to create a "trusted, virtual, federated environment in Europe to store, share and re-use research data across borders and scientific disciplines" (European Commission, 2020).

This research-in-progress explores two research questions: how does the fragmentation of the cloud services landscape affect the pursuit of FAIR data? And, how will pursuing FAIR enable new collaborative workflows? To answer these questions, we look into the Science Mesh as a lens to the wider challenges and promises of FAIR data.

2 Data

This article’s content is based on data produced during the formulation and implementation of the Science Mesh project. Project documentation includes the project proposal containing the rationale, objectives, deliverables, and risk management plan for the project. We also examined the various early materials disseminated by the team to external stakeholders, including those in the project team’s website (cs3mesh4eosc.eu) and articles in partners’ websites and other relevant third party documents. Third, we also had access to the internal wiki by the project team. In this wiki, the project roadmap and all the updates of the various work package teams are well documented. Fourth, we attended the regular project meetings for both the entire project team as well as specific work packages. In these monthly meetings, each work package leader and individual partners give their updates and concerns. Fifth, we also attended meetings where relevant external partners were present. For instance, the Cloud Storage Services for Synchronization and Sharing (CS3) community had a meeting in January 2020 in Copenhagen, Denmark and January 2021 virtually. Finally, we also conducted semi-structured interviews with key members of the team, talking about its goals, similar initiatives, and challenges.

Initially, we relied on the various textual materials and the insights we obtained during project meetings (see Table 1). We then organized these various ideas according to whether they were related to the fragmentation of the cloud landscape, the applications of the cloud in scientific computing, the nature of the mesh or the challenges related to the mesh. After this, we followed up with interviews with key project members to ensure we had a correct understanding of the various components of the mesh. We then iterated on the text, going back to the team frequently for feedback.

Type	Title	Date/s (in 2020)	Pages / Duration
Text	Project Proposal		62p
Text	Internal Wiki	Feb	
Text	Project Website	Feb	5p
Text	Project Blog	Feb	7p
Text	Article in Surf.NL	Mar 2	1p
Text	EOSC Materials		
Meeting	CS3 Conference (Copenhagen, DK)	Jan 27-29	3d
Meeting	CS3 Mesh Kick-Off (Copenhagen, DK)	Jan 30-31	2d
Meeting	Project-wide meetings	Mar 9, Apr 6, May 4, Jun 8, Jul 6	90m - 2h
Meeting	Dissemination work package	Mar 20, Apr 3, Apr 24, May 15, May 29, Jun 12, Jun 19	45m - 1h
Text	Minutes of work package meetings		
Meeting	Open Cloud Mesh	Jun 24	2h
Interview	Project leader	Jun 18	1h
Interview	Work package leader	Jun 23	1h 30m
Interview	Dissemination	Jun 17	1h
Interview	Application	Jul 10	30m
Interview	Application	Jul 14	30m
Interview	Application	Jul 15	20m
Interview	EOSC proponent	Jun 30	30m

Table 1. *Data sources*

3 FAIR data and Cloud Computing

3.1 Promises of FAIR data

Cloud computing has enabled researchers to leverage the advances in computing power and storage capacity together with the decreasing hardware cost (Foster *et al.*, 2008). It has enabled users to deal with the exponential growth in data generated from instrumentation, simulations and archiving. Aside from addressing the computational needs of individual researchers, the cloud also has impacted the way research is conducted worldwide (Jankowski, 2007; Chen and Zhang, 2014; Yang *et al.*, 2014). It has enabled collaborations on increasingly large scales, bringing globally distributed scientists together to work collectively on large research programs. Researchers can assemble different datasets from various sources and access vast computational resources from the convenience of their laptop. Through these collaborations, scientists can pursue previously unimagined lines of scientific inquiry, referred to as e-science. As an example of this transformation in science, Burgelman *et al.* (2019) document how the Ebola and Zika epidemics were addressed quickly by researchers sharing datasets to the public to access. This openness facilitated the quick development of an experimental vaccine.

Moreover, the cloud has been lauded for its potential to reduce waste from unused research data (Tenopir *et al.*, 2015; Madduri *et al.*, 2019). Typically, when researchers finish their projects, they store their data in silos, never to be used again. Yet, other researchers can find new uses for this data that were previously unthought-of when they were first collected by the original researchers. Besides, the advances in distributed machine learning can be used to mine new insights from these datasets.

However, these promises are can only be fully met if data follow FAIR principles. FAIR refers to findable, accessible, interoperable and reusable (Wilkinson *et al.*, 2016; Mons *et al.*, 2017). First, findable means that data can easily be discovered by humans and computers. This involves having clear descriptions of the data, also known as metadata, indexed in an easily searchable registry. Second, data should be accessible through standardised communications protocols. Third, data needs to be interoperable in that they can be easily integrated with other data sets. To make this possible, they should be in a format that can be easily used by other applications. Finally, data should be reusable, following data usage license, provenance, and community standards. Achieving data FAIRness reduces the friction facing international collaborations, opens potentially new frontiers in the conduct of science and reduces waste. FAIR data is a necessary step towards achieving open science (European Commission, 2020), which refers to the practice of science based on cooperative work, promoting transparency and accountability through digital technologies and collaborative tools.

The COVID-19 pandemic has highlighted the consequences of not adhering to data FAIRness. In a letter online, researchers have elaborated the challenges caused by the lack of clear descriptions of primary data generated by different research groups (Schriml *et al.*, 2020). By not having or following the standards in metadata tagging, the data produced by the community has limited reusability, leading to a less-than-optimal response to the pandemic.

3.2 Barriers to FAIR data

Achieving data FAIRness is a complex, multidimensional problem due to social, cultural, institutional, legal, economic, political and technological barriers (Lilleoere and Hansen, 2011; Van Panhuis *et al.*, 2014; Geneviève *et al.*, 2019). On a technical level, the lack of tools and standards can make it difficult to generate and label data of good quality. Ethical and legal considerations may lead to researchers not making their data accessible based on notions of security, privacy and data protection. Political and sociocultural barriers can sow mistrust, leading researchers to be overprotective of their data. In addition, the lack of funding can hinder researchers from allocating time and resources towards data FAIRness activities, away from their primary research. Finally, on an individual level, researchers may have misaligned incentives and motivation against sharing their data. To achieve data FAIRness then, stakeholders have approached it from multiple perspectives. In this article, we focus on an important barrier – the fragmentation of the cloud ecosystem (Ranjan, 2014; Aarestrup *et al.*, 2020).

In the past, the proliferation of public commercial providers such as Amazon, Dropbox, Google and Microsoft have enabled research institutes to cater to their individual users' needs. They have become ubiquitous for providing a wide range of services over the Internet through a pay-as-you-go model. However, as these became widespread, concerns grew especially in the research community whether they can adequately address issues like control, privacy, and data ownership (Mościcki and Mascetti, 2018). As an alternative, on-premise cloud services started to emerge through vendors like NextCloud, OwnCloud and Seafile. These providers are based on the open-source model and can easily be deployed. These allowed research institutes to set up their cloud services easily for researchers.

The proliferation of both private and public cloud storage services however has also caused fragmentation in the ecosystem due to the lack of interoperability across these different options. Cloud interoperability is defined as the ability to seamlessly deploy, migrate and manage application workloads across different hardware and software resources from cloud providers (Ranjan, 2014). Without interoperability, organizations cannot easily exchange or share information and use functionalities across them easily. Moreover, when organizations depend too much on an individual provider, they are prone to vendor lock-in where they cannot easily move to another provider due to high switching costs (Kurze *et al.*, 2011). This is aggravated by network effects. When a certain research institute chooses to host their data on one of these cloud services, they implicitly force users who will interact with their data to use that public cloud service (Vance *et al.*, 2019). These cloud providers then can use these data to improve their service. With users' data attracting more users, it becomes difficult for other companies with smaller user bases to compete In sum, cloud providers are incentivized to put up a walled garden to attract more users and leverage natural monopoly effects.

3.3 Workarounds to FAIR data

To solve this, cloud service providers can agree on common standards to let users easily use data across different environments. However, without much incentive for cloud providers to pursue interoperability, stakeholders from the scientific community have pushed various top-down and bottom-up initiatives to circumvent fragmentation in the pursuit of data FAIRness.

For individual researchers, the common workaround to collaborating across different providers is to move data in and out of the cloud on their own effort. Researchers may even use tools such as container and virtualization technologies to assist with these processes. On a larger scale, another common solution is to create a new, standalone infrastructure where researchers' data and analysis tools are collocated to be readily accessible. This approach is especially deployed in international research collaborations, where a new cloud service is set up to enable remote project teams to store, pool and analyse data. Initiatives such as the Open Science Data Cloud (Grossman *et al.*, 2010) and EUDAT (Lecarpentier *et al.*, 2013) take this idea to a wider scale, aiming to serve the scientific community at large. However, the success of new data commons like these two relies on attracting a critical mass of researchers who would then be able to share data with other users of the service.

As these examples show, the fragmentation problem has been tackled at different levels by the scientific community. However, these solutions have mostly directed the responsibility of FAIR data to the individual researchers and project teams. While these solutions can work on small-scale collaborations, more scalable and robust solutions are desirable to drive the scientific community towards FAIRness. In the following, we discuss an ambitious initiative to tackle the fragmentation challenge, the Science Mesh. We compare it to these previous approaches and from this investigation, we shed light on the challenges and opportunities faced by FAIR data initiatives.

4 Science Mesh

The Science Mesh is a digital infrastructure aiming to enable interoperability across different cloud service providers. In its pilot phase, it will enable frictionless sharing of data across researchers under the following existing services: SURFdrive (Netherlands), CERNBox (Global - CERN), PSNCBox (Poland), CloudStor (Australia), Sciebo (North Rhine-Westphalia, Germany), CESNET (Czech Republic), SWITCHdrive (Switzerland) and ScienceData (Denmark). These services currently host 300,000+ user accounts with over 8.7 PB of storage, equivalent to 2.47 billion files and directories.

For its initial implementation, the mesh has three main components. First, its core component is the interoperability platform which enables the different cloud services to communicate with one another. It extends previous initiatives by the Cloud Storage Services for Synchronization and Sharing community (Mościcki and Mascetti, 2018) to improve interoperability across the various vendors and providers. Second, on top of this platform, the mesh integrates apps like computational notebooks, metadata management and office software, extending their interoperability. Finally, to enable researchers to grant permissions to their collaborators, the mesh uses authentication services (e.g. EduGAIN, 2018) which researchers already use to establish their identity across research institutes.

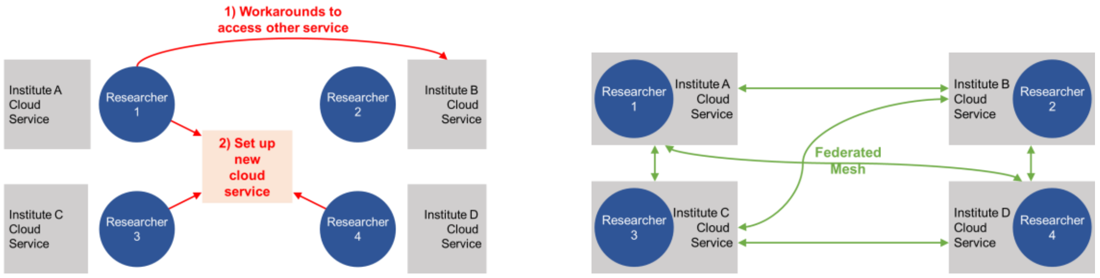
Comparing it to previous workarounds, the Science Mesh aims to directly address the fragmentation challenge by enabling interoperability across cloud providers (see Figure 1). Through this federated approach, users can readily access resources from other cloud providers, without moving outside of their own provider. Users can share their data conveniently with peers without requiring additional steps. From the users' perspective, they do not even need to know that they are working with other cloud providers. By shifting partly the burden of interoperability away from researchers and towards their cloud service providers, the mesh promises to make it convenient to pursue data FAIRness.

4.1 New applications

By enabling interoperability among cloud providers, the mesh promises to unlock new ways of collaborating among researchers. These functionalities within the mesh provide a glimpse of what can also be achieved should data become FAIR across the scientific community.

4.1.1 Frictionless sync and share

At a basic level, the mesh would enable teams to easily share and transfer data between different research sites. This sharing functionality allows collaborators to browse and synchronize files and folders of their peers (like Dropbox). When it is necessary to transfer large datasets, the mesh would facilitate efficient, high-speed transfers.


1) Workarounds to access other service
Institute A Cloud Service
Researcher 1
Researcher 2
Institute B Cloud Service
2) Set up new cloud service
Institute C Cloud Service
Researcher 3
Researcher 4
Institute D Cloud Service
Researcher 1
Institute A Cloud Service
Institute B Cloud Service
Researcher 2
Federated Mesh
Researcher 3
Institute C Cloud Service
Institute D Cloud Service
Researcher 4

Current fragmented landscape

Federated mesh

Figure 1. Approaches to solving fragmentation.

An early adopter for this use case is in astrophysics. With the increasing costs of traditional dish antenna design, there has been a move towards using multiple antennas scattered over a large geographic area. This arrangement, however, produces large datasets generating tens of terabytes every second. This data needs to be moved from the antennas and later combined for processing. In the Low-Frequency Array (LOFAR), for instance, this involves coordination between different storage locations and compute locations (Van Haarlem *et al.*, 2013). The Science Mesh will facilitate the transfer of such large amounts of data through the user-friendly interface of researchers' sync and storage providers.

4.1.2 Remote data analysis

Through the mesh, researchers would be able to interactively analyse large datasets located at a remote site. Researchers would not have to transfer data but instead, carry out all the analysis through the web via access to the data science environment. The key enabling technology are Jupyter notebooks (Kluyver *et al.*, 2016). These notebooks enable researchers to write and run code and create interactive visualizations - all in the same web-based computational environment. These notebooks allow researchers to easily collaborate with peers in prototyping and further developing algorithms. These notebooks also facilitate reproducible research since every step of data analysis can be transparently documented. Through this transparent reporting of scientific procedure, researchers can also modify the code of others for other applications.

An early adopter of this is the Earth Observation community. At the EU Joint Research Center (JRC), researchers apply algorithms to satellite images to detect deforestation in regions around the world (jeodpp.jrc.ec.europa.eu). However, this also requires the collaboration of local partners who can edit and process the input images. Transferring large amounts of data may not be the most appropriate nor efficient for all applications. For instance, to analyse the region of Kenya, it would require around 6 terabytes of image data. To facilitate collaborations then with their partners from Africa, researchers in JRC can embed these satellite images in Jupyter notebooks. These notebooks can then be shared to local partners, enabling them to process the data without requiring data transfer. These research teams can then collaboratively prototype and further develop the code in the notebook.

This capability will also be tested by the high-energy physics community. At the Large Hadron Collider at the European Organization for Nuclear Research (CERN), researchers process the large amounts of data from their experiments by reducing the dataset through elaborate filtering algorithms, then later performing the analysis on this reduced dataset on their local computers. The Science Mesh, however, would allow physicists to analyse much larger datasets by leveraging the cloud infrastructure and tools in big data analysis. One of CERN's experiments showed that 4.7TB of data can be analysed in such a manner (Avati *et al.*, 2019). Through the mesh, researchers can hand access to the computational notebook to their collaborators abroad so that they can perform data analysis on CERN's robust computing platforms.

4.1.3 Collaborative applications

Finally, the mesh will enable collaborators to use apps like document editing and metadata handling software in a federated manner. On one hand, the mesh enables researchers to collaboratively write and edit documents. They would be able to comment and track revisions, like existing services such as Google Docs. In this case, however, the federated mesh allows the files to not be in a centralized repository, giving scientists more control of their data. Moreover, the mesh helps decouple data storage from the applications, enabling scientists much more flexibility on their preferred software to analyze their data.

Second, the mesh would enable researchers to collaboratively label and add metadata to datasets and publish them with persistent identifiers directly through the Science Mesh. By enabling full metadata awareness in the research workflow, this ultimately enables the reuse of data. These functionalities are supported by integrating tools developed by OpenAIRE (Rettberg and Schmidt, 2012) for open dataset publication. As funding sources increasingly demand that researchers publish data used in academic articles, the Science Mesh will allow researchers to publish their data into open repositories from the convenience of the interface of their cloud provider.

Principle	Description	Current setting	Federated mesh
Findable	Metadata and data are easy to find for both humans and computers.	Data may not be properly labelled for easy search	Also subject to existing barriers of data not being properly labeled
Accessible	Once data is found, they can be accessed with authorisation and authentication	Data may be stored in silos that are difficult to access	Data can be accessed through existing authentication systems
Interoperable	Data can be readily used by applications for analysis, storage, and processing.	Workarounds to use data and apps outside their original service	Data and apps can be used across systems through interoperability layer
Reusable	Metadata and data are well-described for replication and integration	Data is stored or archived without considering reusability	Tools towards reusability are integrated into the workflow

Table 2. Comparison of cloud setups in enabling FAIR principles (Adapted from go-fair.org)

This functionality will be especially useful in managing archival collections. For instance, the Pacific Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) is a digital archive, containing records of many small cultures and languages around the world (Thieberger and Barwick, 2012). They store various items related to and recordings of over 1,200 languages, amounting to 84 Tb. To effectively process and archive such data, researchers must be able to collaborate on labelling and processing such files in their catalogue. Once processed and published, these files can then be accessed by other researchers for further analysis through the mesh.

4.2 Challenges

The success of the Science Mesh hinges on gaining the support of two groups of stakeholders: cloud service providers and the research community. On one hand, to implement the Science Mesh, cloud vendors upstream would have to open their service sufficiently such that they can plug into the mesh and be interoperable with other vendors. In the first phase of the project, three vendors OwnCloud, NextCloud and Seafile, which serve the eight previously mentioned cloud services, would be included in the mesh. The hope is that by achieving a critical mass, the mesh can convince other providers such as Dropbox, Amazon, Microsoft and Google to also be part of the mesh in the future. Further research is necessary to make the business case for cloud vendors to connect to the mesh

Second, the long-term sustainability of the mesh ultimately depends on whether researchers use the service. By building on top of services that researchers already use, the initiative lowers the adoption barriers as researchers would only have to use their institute’s existing service to collaborate with external parties. To further attract researchers to use the mesh, the advanced functionalities mentioned before could be crucial. Nonetheless, further research is needed to evaluate how the mesh could overcome the social, cultural and institutional barriers to data sharing.

5 Conclusions and Next Steps

This research-in-progress explores the challenges and potential of FAIR data. Using the Science Mesh initiative as a lens reveals that FAIR data is a social dilemma acting on two levels. On one hand, making data FAIR can be costly for individual researchers, requiring them to use various workarounds to make their data interoperable across different providers. On the other hand, cloud service providers are not incentivized to coordinate with one another to achieve interoperability. These two dilemmas interact with one another: The lack of cloud interoperability makes it difficult for researchers to make their data FAIR. Without much demand from users for FAIR data, cloud providers then are not compelled to make their services interoperable.

We then looked into the promises of the Science Mesh in connecting the fragmented cloud landscape. On its first phase, the mesh will link eight cloud providers serving 300,000 researchers. From this initial implementation, it hopes to build a critical mass to then compel other cloud providers to open their walled gardens and plug to the mesh. This initial implementation would then be crucial in informing the future challenges as the mesh expands.

The mesh aims to make it convenient for researchers to make their data FAIR by federating their existing cloud provider and attracting them through advanced functionalities such as frictionless sync and share, remote data analysis and collaborative applications. We provided examples of the potential of FAIR data by enabling novel research workflows in fields including astronomy, high energy physics, earth science and the humanities. Moving forward, further empirical studies will be conducted to explore the challenges of FAIR data in these lead user communities. Case studies will also be conducted to explore how these advanced functionalities can fuel novel ways of conducting science.

As the volume of data generated by the research community accelerates, reducing the inconveniences in data sharing will be crucial to ensure that no data goes to waste and that data is used effectively to inform important research questions. An important step to FAIRness is ensuring that data and the tools used to analyze them can operate across different cloud providers. If successful, the Science Mesh will play an important role in enabling collaborative research towards solving the increasingly complex problems facing society.

Acknowledgments

This project was funded by the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement No. 863353 CS3MESH4EOSC.

References

Aarestrup, F. M. *et al.* (2020) 'Towards a European health research and innovation cloud (HRIC)', *Genome Medicine*. Genome Medicine, 12(1), p. 18. doi: 10.1186/s13073-020-0713-z.

Avati, V. *et al.* (2019) 'Declarative Big Data Analysis for High-Energy Physics: TOTEM Use Case', in, pp. 241–255. doi: 10.1007/978-3-030-29400-7_18.

Benfeldt, O., Persson, J. S. and Madsen, S. (2020) 'Data Governance as a Collective Action Problem', *Information Systems Frontiers*. doi: 10.1007/s10796-019-09923-z.

Berriman, G. B. *et al.* (2013) 'The application of cloud computing to scientific workflows: A study of cost and performance', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1983). doi: 10.1098/rsta.2012.0066.

Burgelman, J.-C. *et al.* (2019) 'Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century', *Frontiers in Big Data*. doi: 10.3389/fdata.2019.00043.

EduGAIN (2018) *What is EduGAIN*. doi: https://edugain.org/about-edugain/what-is-edugain/.

European Commission (2019) *Cost-benefit analysis for FAIR research data*.

European Commission (2020) *European Open Science Cloud (EOSC) Partnership*.

Foster, I. *et al.* (2008) 'Cloud Computing and Grid Computing 360-Degree Compared', in *2008 Grid Computing Environments Workshop*. IEEE, pp. 1–10. doi: 10.1109/GCE.2008.4738445.

Geneviève, L. D. *et al.* (2019) 'Factors influencing harmonized health data collection, sharing and linkage in Denmark and Switzerland: A systematic review', *PLoS ONE*. doi: 10.1371/journal.pone.0226015.

Grossman, R. L. *et al.* (2010) 'An overview of the Open Science Data Cloud', in *HPDC 2010 - Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*. doi: 10.1145/1851476.1851533.

Van Haarlem, M. P. *et al.* (2013) 'LOFAR: The low-frequency array', *Astronomy and Astrophysics*. doi: 10.1051/0004-6361/201220873.

Hoffa, C. *et al.* (2008) 'On the use of cloud computing for scientific workflows', *Proceedings - 4th IEEE International Conference on eScience, eScience 2008,* pp. 640–645. doi: 10.1109/eScience.2008.167.

Jankowski, N. W. (2007) 'Exploring e-science: An introduction', *Journal of Computer-Mediated Communication*. doi: 10.1111/j.1083-6101.2007.00337.x.

Kluyver, T. *et al.* (2016) 'Jupyter Notebooks—a publishing format for reproducible computational workflows', in *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016*. doi: 10.3233/978-1-61499-649-1-87.

Kurze, T. *et al.* (2011) 'Cloud Federation', *Computing*, (c), p. 7.

Lecarpentier, D. *et al.* (2013) 'EUDAT: A New Cross-Disciplinary Data Infrastructure for Science', *International Journal of Digital Curation*. doi: 10.2218/ijdc.v8i1.260.

Lewis, G. A. (2013) 'Role of standards in cloud-computing interoperability', in *Proceedings of the Annual Hawaii International Conference on System Sciences*. doi: 10.1109/HICSS.2013.470.

Lilleoere, A. M. and Hansen, E. H. (2011) 'Knowledge-sharing enablers and barriers in pharmaceutical research and development', *Journal of Knowledge Management*. doi: 10.1108/13673271111108693.

Madduri, R. *et al.* (2019) 'Reproducible big data science: A case study in continuous FAIRness', *PLoS ONE*, 14(4), pp. 1–22. doi: 10.1371/journal.pone.0213013.

Mendez, K. M. *et al.* (2019) 'Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing', *Metabolomics*. Springer US, 15(10), p. 125. doi: 10.1007/s11306-019-1588-0.

Mons, B. *et al.* (2017) 'Cloudy, increasingly FAIR; Revisiting the FAIR Data guiding principles for the European Open Science Cloud', *Information Services and Use*. doi: 10.3233/ISU-170824.

Mościcki, J. T. and Mascetti, L. (2018) 'Cloud storage services for file synchronization and sharing in science, education and research', *Future Generation Computer Systems*. doi: 10.1016/j.future.2017.09.019.

Ortiz, S. (2011) ‘The problem with cloud-computing standardization’, *Computer*. doi: 10.1109/MC.2011.220.

Ostrom, E. (1990) *Governing the Commons, Governing the Commons*.

Van Panhuis, W. G. *et al.* (2014) ‘A systematic review of barriers to data sharing in public health’, *BMC Public Health*. doi: 10.1186/1471-2458-14-1144.

Philip Chen, C. L. and Zhang, C. Y. (2014) ‘Data-intensive applications, challenges, techniques and technologies: A survey on Big Data’, *Information Sciences*. Elsevier Inc., 275, pp. 314–347. doi: 10.1016/j.ins.2014.01.015.

Ranjan, R. (2014) ‘The Cloud Interoperability Challenge’, *IEEE Cloud Computing*, 1(2), pp. 20–24. doi: 10.1109/MCC.2014.41.

Rettberg, N. and Schmidt, B. (2012) ‘OpenAIRE — Building a collaborative open access infrastructure for european researchers’, *LIBER Quarterly*. doi: 10.18352/lq.8110.

Schriml, L. M. *et al.* (2020) ‘COVID-19 pandemic reveals the peril of ignoring metadata standards’, *Scientific Data*. doi: 10.1038/s41597-020-0524-5.

Tenopir, C. *et al.* (2015) ‘Changes in data sharing and data reuse practices and perceptions among scientists worldwide’, *PLoS ONE*. doi: 10.1371/journal.pone.0134826.

Thieberger, N. and Barwick, L. (2012) ‘Keeping records of language diversity in melanesia: the pacific and regional archive for digital sources in endangered cultures (PARADISEC)’, *Melanesian languages on the edge of Asia: Challenges for the 21st Century*, 5, pp. 239–253.

Vance, T. C. *et al.* (2019) ‘From the Oceans to the Cloud: Opportunities and challenges for data, models, computation and workflows’, *Frontiers in Marine Science*, 6(APR), pp. 1–18. doi: 10.3389/fmars.2019.00211.

Wilkinson, M. D. *et al.* (2016) ‘The FAIR Guiding Principles for scientific data management and stewardship’, *Scientific Data*, 3(1), p. 160018. doi: 10.1038/sdata.2016.18.

Wise, J. *et al.* (2019) ‘Implementation and relevance of FAIR data principles in biopharmaceutical R&amp;D’, *Drug Discovery Today*, 24(4), pp. 933–938. doi: 10.1016/j.drudis.2019.01.008.

Yang, X. *et al.* (2014) ‘Cloud computing in e-Science: Research challenges and opportunities’, *Journal of Supercomputing*, 70(1), pp. 408–464. doi: 10.1007/s11227-014-1251-5.