SCIENCE|BUSINESS®

The European Open Science Cloud: Who pays for what?

A report of the Science|Business Network's Cloud Consultation Group

February 2018

On 4 December 2017, members of the Science|Business Network's Cloud Consultation Group met to discuss the economics of the European Union's proposed "science cloud" project. This report is based partly on those discussions, but is ultimately a product of Science|Business. The views expressed herein do not necessarily reflect those of individual members.

This report is the third in a series that the group is working on, to gather private and public sector expertise on topics of importance to the development of the European Open Science Cloud. The other reports are available on www.sciencebusiness.net:

The case for the cloud. May 2017
Governing the Open Science Cloud. October 2017

Cloud Consultation Group members:

Amazon
Association for Computing Machinery – Europe Policy Committee
Association of Commonwealth Universities
Barcelona Supercomputing Centre
CERN
ETH-Zurich
European University Association (EUA)
European Space Agency
GÉANT
Google
Huawei
Microsoft
Research Data Alliance Europe
University of Eastern Finland
University of Twente

Rapporteur: David Pringle

Avenue des Nerviens 79, Brussels, Belgium 1040
info@sciencebusiness.net
+322-304-7577

Executive summary

Drawing on the work of the Science|Business Network's Cloud Consultation Group, this report explores the economics of the European Open Science Cloud (EOSC) – an ambitious EU initiative to enable Europe's 1.7 million researchers to build on each other's research and accelerate scientific progress.

The costs of building the EOSC

Although the EOSC will initially harness the existing digital infrastructures used by European science, it will require further investment. There are essentially seven significant, but partially overlapping, categories of cost associated with the EOSC:

1. Employing cloud-computing services: The cost of getting data into the cloud and storing some of it for decades, and the cost of using cloud computing resources to access and analyse scientific data, including the necessary connectivity.
2. Opening up scientific data: The implementation of data management plans to make research data findable, accessible, interoperable and re-usable (FAIR principles).
3. Federation of existing scientific data infrastructures with new provisioning schemes, such as the cloud or specialised facilities, and the development of nodes to link existing national data centres, European e-infrastructures, external providers and research infrastructures.
4. Development of specifications for application interoperability (APIs), data portability and data sharing: To enable data to be shared across disciplines and infrastructures, more standardisation of meta-data and, perhaps, the actual data itself will be needed.
5. Creation of search tools: New software will be required to enable scientists to search, browse and access research data.
6. Creation and maintenance of a secure environment: The European Commission envisions a suitable certification scheme will be designed at EU level to guarantee security, data portability, and interoperability in compliance with legal requirements. Such a scheme will need to be flexible enough to enable the EOSC to keep pace with the evolution of scientific research.
7. The governance of the EOSC process: The EOSC will need a full time executive body that can oversee federation, long-term funding, sustainability, data preservation and stewardship. See the consultation group's report, Governing the European Open Science Cloud, for recommendations on how the EOSC could be run.

The economic benefits of the EOSC

By putting cutting-edge computing resources at the fingertips of researchers, the open science cloud could bring about a step change in productivity. The availability of computing resources should no longer be a bottleneck. If, as commercial cloud providers say, only a small percentage of European science is taking advantage of so-called hyper-scale cloud technologies today, there is enormous scope for a transformation in the way in which researchers share and analyse data. The implementation of the EOSC could catalyse widespread adoption of hyper-scale cloud computing by European science.

Ultimately, the EOSC could have a profound impact on European scientists' capabilities, giving them access to a multitude of platforms, software tools, algorithms and data that they can't access today. By creating a safe and seamless environment for sharing research data, the EOSC could bring about dramatic change in scientists' productivity. As a result, researchers in both the public and private sectors will be able to conduct new kinds of experiments and research, with a lower level of risk, which could ultimately yield major economic benefits. The net effect would be to breathe new life into existing investments and draw new money into European science, creating a virtuous circle that fuels investment in innovative businesses and new public services. See the group's report, The Case for the Cloud, for more on how the EOSC could transform European science.

Funding the EOSC

The European Commission has allocated €260 million for the federation of the existing scientific data infrastructures, and has promised EU member-states that the EOSC will be self-sustaining after 2020. In theory, this sum could be supplemented by a further €12 billion per annum, made up of the approximately €10 billion a year already spent on data infrastructures for science conducted in European universities and other publicly-owned facilities, plus 1% of the €200 billion a year of public money spent on scientific research. Although the EOSC's high level expert group estimated up to 5% of research budgets eventually may have to be allocated to data management, the Science|Business group believes 1% to 2% may be sufficient, once the EOSC has matured. However, as the benefits become apparent, some institutions may invest a larger slice of their budgets in data management.

Research funding agencies could make a small percentage of each grant available as credits that can be spent on any kind of cloud service (so long as it meets the EOSC's technical/security/privacy criteria). This approach would help drive competition between cloud providers'; the researcher's IT specialists would spend the credits with the provider offering the best value. Although research funders should insist that grantees make their data open and compatible with the EOSC, the grants should be agnostic about what cloud services they use to make their data findable, accessible, interoperable and reusable.

To maximise the effectiveness of the money spent on the EOSC, investments in the initiative should be driven by demand, rather than a "build it and they will come" mentality. Demand is likely to be particularly strong for platform-as-a-service capabilities, which can help to significantly reduce the effort required to develop the algorithms and software researchers need for their projects. Where possible, depending on the scientific discipline, the EOSC should not require data to be transferred from one place to another; it is more efficient to store data in a single location and perform analytics in that location, rather than create multiple copies of a large data set.

Monetising the EOSC

Over time, the EOSC could also generate its own income stream by serving the needs of the private sector. Although the EOSC is intended to make research data free at the point of use for scientists, commercial entities could be required to pay to access data within the EOSC framework once their usage rises beyond a specific threshold. A points-based application process, designed to gauge the public value of the project in the broadest sense, could be used to determine the thresholds that apply to each entity.

However, there are many other ways in which the EOSC could be monetised, so the business model will need to be carefully conceived and refined over time. This will be the subject of a future report.

Ultimately, the data within the EOSC could underpin an ecosystem of commercial services, just as, today, the satellite data captured by the European Space Agency is being used as the basis for commercial offerings. Given the value that the EOSC could bring to private sector research and product development, it could potentially build up a substantial revenue stream over time.

However, another school of thought argues that the EOSC may not need to generate any revenues, as it will become self-sustaining in the same way that open source software is maintained by its community of users (typically with some support from large technology companies). In this scenario, individual researchers, empowered to employ whichever platform makes most sense to them, will then be doing nearly all their work using publicly developed and widely shared mobile workloads. As scientists re-use and enhance each other's workloads, they will be improving and expanding the EOSC, which will take on a life of its own akin to that of the open source software movement.

Introduction

How should Europe fund its lofty plans for an open science cloud? A far-reaching and multi-faceted undertaking, the European Open Science Cloud (EOSC) initiative aims to provide Europe's 1.7 million researchers and 70 million students and professionals in science and technology with easy access to other researchers' data, and to a wide range of computing resources. What's more, the Commission has said, the data in this virtual environment is supposed to be "free at the point of use."

The goal is to ensure researchers across the EU can access open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines. To that end, data held in the EOSC will be governed by the so-called FAIR principles – all data needs to be findable, accessible, interoperable and reusable.

But bringing about this open science nirvana will cost money and someone will have to pay. And it won't be the taxpayer: The Commission has promised EU member states that they won't need to find new money and the EOSC will be self-sustaining by 2020.

As yet, no comprehensive cost-benefit analysis of the EOSC exists. That's partly because no one actually knows how much Europe spends on managing scientific data today and partly because it is impossible to anticipate the economic impact of the scientific breakthroughs that may or may not be catalysed by the EOSC. Ultimately, the initiative may need to be underpinned by, as yet unspecified, business models that will enable the science cloud to be self-sustaining.

The architects of the EOSC urgently need to figure out who will pay for what and how much is being spent on the current data infrastructure. Other related questions include: What will the EOSC really mean for the cost of handling data? How will the internal accounting work? What will be the total bill and how will it change over time?

This paper starts to address some of these thorny questions. It begins by outlining how European science is employing the cloud today and the different categories of costs involved in establishing the EOSC. It then identifies the potential efficiency benefits associated with a move to cloud computing and open data sharing, before considering some of the longer-term economic benefits that could arise from an open science cloud. The paper concludes by looking at the potential sources of funding and making some recommendations for the EOSC's many stakeholders.

1. How Europe's researchers are using the cloud today

Europe's scientists are making limited use of cloud technologies today, significantly lagging behind adoption in the private sector. Commercial cloud providers say only a small percentage of European science employs the latest "hyper-scale" cloud services[1] and tools. At the other end of the spectrum, many researchers still share their findings by emailing spread-sheets. In the middle, the majority of researchers make do with a diverse mixture of in-house databases, private clouds and online commercial services.

Today, fragmentation abounds: There is limited integration or interoperability between the information and communications technologies (ICT) used by European scientists, while the longevity and openness of research data varies both across disciplines and within disciplines. Most data generated from scientific research is still kept in isolation and trapped in silos. Although a number of National Research and Education Networks (NRENs) offer cloud services and some large research institutions, such as CERN, have their own well-established cloud services, these are not necessarily inter-connected and are mostly limited to IaaS (infrastructure-as-a-service).

More broadly, Europe's public sector is trailing some way behind the private sector in employing cloud services. In its April 2016 Communication outlining the proposed science cloud, the Commission accurately described the uptake of cloud services in the public sector as "uneven and slow," blaming a lack of trust and limited synergies within the public sector and academia. It also flagged the fragmentation of data infrastructures as an obstacle for building critical mass and common solutions for different user groups.

Indeed, there is a plenty of evidence (see next section) that European scientists are ill equipped to manage and process the vast amount of data being captured by the latest research instruments.

As the pace of innovation intensifies, there is growing economic pressure on European science and research to become more agile and for Europe to develop a much more advanced and sophisticated research infrastructure. Some policymakers worry whether Europe has sufficient computing capacity to meet the fast growing demands of so-called big science, which involves crunching the big data being captured by the increasingly digital economy. In its April 2016 Communication, the Commission warned that Europe lags in terms of sheer total computing power: only one of the ten leading high performance computing (HPC) infrastructures is in the EU, with Germany's Höchstleistungsrechenzentrum Stuttgart ranking 8th. The USA has five; and China has had the fastest supercomputer in the world since 2013.

However, some experts believe a fixation with owning HPC infrastructure is misplaced, arguing that the key infrastructures of 21st century science are datasets, tools and techniques, rather than computing capacity. They contend, that Europe (as a hotbed of global science, see Figure 1) already has much of this "soft infrastructure", but it needs knitting together and to become much more sharable.

1. Including commercial infrastructure-as-a-service offerings procured through the GÉANT framework agreements.

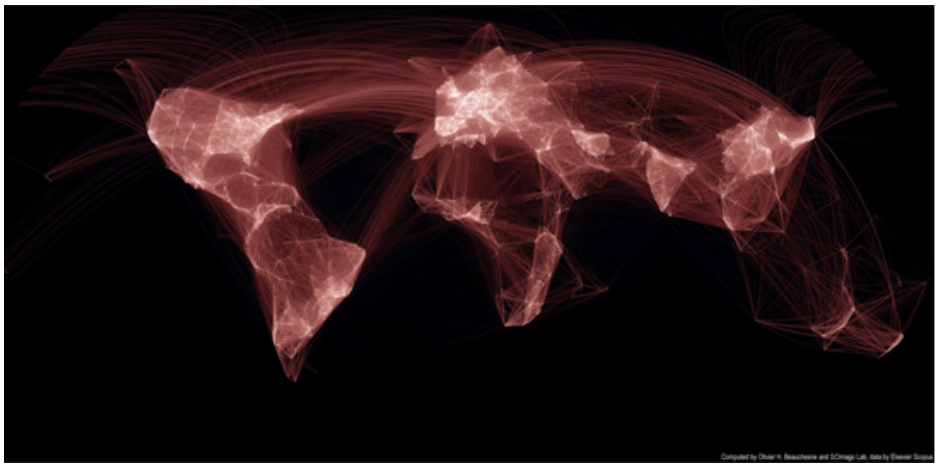

Figure 1 – Map of scientific collaboration between researchers: Europe is a focal point (Source: Olivier H. Beauchesne - http://bit.ly/e9ekP2)

The costs of inefficient data management

An inability to efficiently manage and process data continues to hold back European science, according to a 2016 study by the German Radieschen[1] and two EC-funded projects EUDAT and RDA[2] , based on 50 interviews and 80 intensive discussions with experts from various disciplines and organisations. The study found that data management is either non-existent, incomplete or ends up being a huge drain on resources. One of the biggest issues is an inconsistent approach to organising data, making subsequent data management and processing too time-consuming and costly: Many researchers still use basic file systems.

The study also found that it is so expensive to create logical layer information, which is required to trace provenance, understand creation context, check identity and integrity, that it simply doesn't happen (even though most data professionals understand this hands-off model is unsustainable). Moreover, a haphazard approach to automation is making it difficult or impossible to manage and process so-called big data: Too many ad hoc scripts without proper documentation are used. In a similar vein, a lack of software to support proper data organisations leads to the creation of legacy data that cannot be integrated easily with other data.

1. The Radieschen (Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur) project aims to define a future data infrastructure for research in German.
2. The EUDAT project to develop a collaborative data infrastructure and the Research Data Alliance.

Most disciplines have a major problem with legacy data, yet continue to create it.

Today's data management practices can be a huge waste of money and skills: The study found one of the key biologists in a large research institute is spending 75% of his time on manual data management. At the DAMDID[1] conference in Obninsk in 2015, Michael Brodie from MIT reported the findings of a study in the US that 80% of the scientist time in data-driven projects is wasted with "data wrangling" - the work that needs to be done before the real analysis can start. At the Big Data Summit 2017, industry experts reported about 60% of the costs of data-driven projects are consumed by data wrangling work.

Such estimates are supported by anecdotal evidence. At a cancer research institute in Germany, researchers have not moved to semi-automatic workflows since there are too many exceptions and parameter variations to be considered. Instead, they rely on ad hoc scripts and manual management steps, meaning 75% of time is wasted. The equivalent US institute has hired IT experts for three years to work closely with the researchers and develop a flexible workflow system. It has invested about $600,000 to reduce the wasted time on data management.

1. Data Analytics and Management in Data Intensive Domains conference.

2. Costs of establishing and maintaining the EOSC

The European Open Science Cloud (EOSC) is meant to be a flexible and versatile tool that can be used for many different purposes. As well as enabling scientific research, the EOSC will be open for education and training purposes in higher education and, over time, to government and business users.

The June 2017 EOSC Declaration, which has been endorsed by a wide range of stakeholders, calls for the EOSC to support several different cloud deployment models, including infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS) and software-as-a-service (SaaS), to meet the needs of research communities at different levels of maturity. The EOSC is also intended to support the whole research lifecycle through the provision of a wide set of software, infrastructure, protocols, methods, incentives, training and services.

Still, the development and maintenance of the EOSC will call primarily for investment in software, rather than hardware. Although the Commission is intending to support the deployment of more high performance computing (HPC) capacity, a primary goal is to enable the development of a Europe-wide marketplace for existing scientific data and cloud computing resources, rather than build a new tranche of specialist infrastructure. The June Declaration makes it clear that the EOSC will be a managed marketplace, rather than a dedicated cloud infrastructure controlled top-down by the EU institutions.

Furthermore, the EOSC isn't an isolated initiative. It is part of a concerted effort by the Commission to promote open science in Europe, through open access to the scientific publications and data funded by the EU's Horizon 2020 programme.

Bearing these broad principles in mind, there are essentially seven significant, but partially overlapping, categories of costs associated with the EOSC.

Cost category	Description
Employing cloud-computing services	The cost of getting data into the cloud and storing it for decades, and the cost of using cloud computing resources to access and analyse scientific data, including the necessary connectivity.
Opening up scientific data	The implementation of data management plans to make research data findable, accessible, interoperable and re-usable (FAIR principles).
Federation of existing scientific data infrastructures	Nodes will be needed to link national data centres, European e-infrastructures and research infrastructures
Development of specifications for interoperability and data sharing	To enable data to be shared across disciplines and infrastructures, more standardisation of meta-data and perhaps the data itself will be needed.
Creation of search tools	New software tools will be required to enable scientists to search, browse and access research data.
Creation and maintenance of a secure environment	The European Commission envisions a suitable certification scheme will be designed at EU level to guarantee security, data portability, and interoperability in compliance with legal requirements
The management of the EOSC	The EOSC will need a full time executive body that can oversee federation, long-term funding, sustainability, data preservation and stewardship

Figure 2 – The seven main cost categories for the EOSC

1. Cost of employing cloud-computing services. The EOSC will enable scientists to make much greater use of cloud computing services, which will need to be paid for. There are two dimensions to this. First is the cost of getting data into the cloud and storing it for decades, so future research projects can easily access it. Second is the cost of using cloud computing resources to access and analyse scientific data, including the necessary connectivity.

Although data within the scope of the EOSC is intended to be free for registered users to access at the point of use, researchers will need to pay to use the cloud computing resources required to analyse and manipulate the data. As this marketplace will be demand-driven, the June 2017 Declaration envisions that both public research data Infrastructures and commercial operators will develop and provide services based on user needs, and discontinue provision when not justified by the level of adoption.

As the EOSC cloud services will be provided in a competitive marketplace by any compliant supplier, service providers will aim to make the cost of using this cloud-based market attractive to researchers relying on the in-house IT being used today. These costs could be met in part by a reallocation of existing e-infrastructure budgets, supplemented by a small percentage of research budgets. The EOSC could also generate some of the revenue from commercial users.

2. Cost of opening up scientific data: Academics, industry and public services will have to be persuaded and incentivised to share their data, and improve their data management training, literacy and stewardship skills. In some cases, they will have no choice – the provision of open data will be made a condition of research grants. The Commission has said that all the scientific data produced by the Horizon 2020 programme will be open by default. Research projects covered by the programme will have to implement data management plans to make research data findable, accessible, interoperable and re-usable (FAIR principles).

The high level expert group advising the Commission on the EOSC has recommended that well-budgeted data stewardship plans should be made mandatory and about 5% of research expenditure should be spent on properly managing and stewarding data. However, some experts believe this figure is unnecessarily high – 1-2% of research budgets may be sufficient once the EOSC matures. However, as the benefits become apparent, some institutions may invest a larger slice of their budgets in data management.

3. Federation of existing scientific data infrastructures, which are scattered across disciplines and Member States and are innovating at different speeds.New provisioning schemes, such as the cloud or specialised facilities, and nodes will be needed to link national data centres, European e-infrastructures and research infrastructures. This federation will be a critical building block for the EOSC. The June 2017 Declaration calls on service provision to be based on local-to-central subsidiarity (e.g. national and disciplinary nodes connected to nodes at a pan-European level). It envisions that data will be progressively federated by the creation of open data infrastructures developed in specific thematic areas (e.g. health, environment, food, marine, social sciences, transport). This could be a major undertaking: a lot of existing research data may have to be moved or processed and analysed at considerable expense.

The European Commission has allocated €260 million to pay for the initial federation work. This budget could be used to run multiple tenders in which different working groups/consortia conduct experiments to find out which approaches work better than others. However, there will also be other calls on this seed budget, which may be insufficient to stitch together the very diverse infrastructure provided by the private and public sectors. Ultimately, this work may need to be completed using revenue generated by commercial users of the EOSC.

4. Development of specifications for interoperability and data sharing across disciplines and infrastructures, building on existing initiatives, such as the Research Data Alliance and the Belmont Forum, and legal provisions, such as INSPIRE, and the GO FAIR Movement. While interoperability and data sharing are already being tackled in some sectors (e.g. location of data by the INSPIRE Directive, health data by the Patients' Rights Directive), many data sets remain unavailable to scientists, industry, public administrations and policymakers. The use of consistent meta-data standards (see cost category 5 for more on this) will help address this problem, but more standardisation of the data sets themselves may also be required.

The Commission anticipates that, over time, the Digital Single Market Priorities for ICT Standardisation will fill many of the gaps in the current specifications. However, some experts believe standardisation of data sets may prove too difficult, as some researchers will want to reserve the right to use their own distinctive approaches to data management. In any case, the implementation of interoperability should be demand-driven – there is demand for certain classes of interoperability, but not for all data interoperability.

Most of this work will be financed by existing standardisation initiatives, but it may also need to draw on revenue generated by commercial users of the EOSC (see chapter 5).

5. Creation of search tools: The development of new software tools will be required to enable scientists to search, browse and access research data. These will make use of the meta-data that annotates and identifies the underlying data. This meta-data will need to conform with specifications that enable it to be processed through common, open source data analysis tools. Ideally, all research data will be available programmatically, through web APIs, so that it can be identified and accessed by search engines and automated systems.

While the creation of meta-data will probably be funded from research grants (see cost-category 2), search tools could be made available as either commercial or public PaaS cloud services that act as gateways into different data sets. In both cases, the development of the tools will likely be funded by revenue generated by commercial users of the EOSC (see chapter 5).

6. Creation and maintenance of a secure environment where privacy and data protection are guaranteed and users can be confident they won't face data security and liability risks. The European Commission envisions a suitable certification scheme will be designed at EU level to guarantee security, data portability, and interoperability in compliance with legal requirements.

This certification scheme will need to be flexible enough to keep pace with the evolving requirements of European science. There is also a risk that it could lead to the development of a self-interested certification industry that could dilute the impact of the EOSC. The EOSC may also need to secure exemptions from some of the most onerous data protection provisions in the new General Data Protection Regulation, which could hinder data portability and reusability. Although some narrow certification schemes already exist, the Commission notes there is no common EU-wide approach to the procurement or secure management of public sector cloud resources. To manage access to research data and tools, the EOSC will need a robust authentication system, which combines a single sign-on process, resulting in a federated identity and credentials for all users of the EOSC. Clearly, such a system should build on the existing systems employed by the research infrastructures to authenticate members of their respective user communities. To keep costs down, the Commission anticipates the EOSC's security will be based on existing public sector initiatives, such as the Connecting Europe Facility Digital Service Infrastructure building blocks related to trust and security.

Initially, the development of a certification and authentication system may need to be funded by the Commission's €260 million budget and/or the Horizon 2020 budget, but it may also need to tap commercial revenues generated by the EOSC (see chapter 5).

7. Funding the management of the EOSC. The EOSC will need a full time executive body that can oversee federation, long-term funding, sustainability, data preservation and stewardship.

Initially, the management structure will probably need to be funded by the Commission's €260 million federation budget, but it will also need to draw on revenue generated by commercial users of the EOSC (see chapter 5).

High performance computing and the European Data Infrastructure

The original EOSC plan calls on Europe to build an integrated world-class high performance computing (HPC) capability, high-speed connectivity and leading-edge data and software services. In the April 2016 Communication, the Commission called for the implementation of a "European Data Infrastructure" that will employ exascale supercomputers based on EU technology to lift the EU into the world's top supercomputing powers.

The Commission believes a world-class HPC infrastructure is required to handle the most demanding scientific and engineering use cases, such as simulating a complete next-generation airplane, climate modelling, linking genome to health and understanding the human brain. In March 2017, eight EU member states signed the "European commitment to HPC", agreeing to work together and with the European Commission to acquire and deploy a pan-European integrated exascale supercomputing infrastructure by 2022/2023. Exascale computing systems are capable of one billion billion calculations per second.

The Commission has estimated the EU will need to spend €3.5 billion on data infrastructure and €1 billion on a large-scale EU-wide quantum technologies flagship over a period of five years. It noted that potential sources of EU financing include:

- Horizon 2020 Framework Programme for Research and Innovation (Horizon 2020)
- Connecting Europe Facility (CEF)
- European Structural and Investment Funds (ESIF)
- European Fund for Strategic Investments (EFSI)

Although the planned data infrastructure would clearly support the development of the EOSC, building an HPC capability is not a pre-requisite for the EOSC.

While it will eventually make use of dedicated HPC infrastructure, the EOSC is likely to rely primarily on cloud computing resources supplied by private companies and existing public institutions, rather than the proposed exascale supercomputers. Although government labs and academia are still leading spenders on HPC (with a global spend of $2 billion and $1.9 billion respectively in 2016), commercial spending on HPC is growing rapidly, and now accounts for 52% of the market, according to Hyperion Research[1].

Cloud service providers are investing in HPC to meet rising demand in the private sector for advanced analytics (supported by artificial intelligence) to enable fraud detection, affinity marketing, business intelligence, and precision medicine, as well as data-driven science and engineering, intelligence/security analytics, and knowledge discovery.

Indeed, Europe's research infrastructure is becoming increasingly integrated with commercial cloud computing services. For example, the pan-European research and education network GÉANT is making it straightforward for researchers to use tailor-made cloud computing services from Microsoft and Amazon Web Services via an IaaS framework. The framework is designed to make IaaS resources more readily accessible to the 50 million or so users of GÉANT's pan-European 500Gbps network by removing the need for individual tenders and enabling users to benefit from volume discounts. GÉANT says this demand aggregation model is generating significant savings for the National Research and Education Networks that have signed up providers to the IaaS contracts.

Even so, the EU will still need to invest public funds in HPC. Some specialised research programmes will continue to need their own dedicated infrastructure, which will need to be incorporated into the EOSC.

1. https://www.hpcwire.com/2017/11/15/hyperion-hpc-market-update-decent-growth-led-hpe-ai-transparency-risk-issue/

3. The immediate benefits of the EOSC

By putting cutting-edge computing resources at the fingertips of researchers, the open science cloud could bring about a step change in productivity. In particular, the implementation of the EOSC could catalyse broader adoption of so-called hyper-scale cloud computing by European science. In theory, at least, this shift could generate financial savings that could be used to pay for the costs outlined earlier in this paper, such as compliance with the FAIR principles and the creation of a secure environment.

The Commission's vision is that the EOSC will ensure every research centre, every research project and every researcher in Europe has access to the computing, data storage and analysis capacity they need to conduct the research they want to do; the availability of computing resources should no longer be a bottleneck. If, as experts say, only a small percentage of European science is taking advantage of hyper-scale cloud technologies today, there is enormous scope for a transformation in the way in which researchers share and analyse data. This section considers two sets of immediate benefits for researches that should arise from the implementation of the EOSC:

- The benefits of moving from an in-house IT infrastructure to cloud-based solutions
- The benefits that relate specifically to the implementation of the EOSC

Efficiency gains arising from cloud computing

Traditionally, both private companies and public organisations have designed, deployed and managed their own IT infrastructure, either buying sufficient hardware and software to cope with peaks in demand or capping usage to meet the capacity available. Cloud computing is flexible in that an organisation can rent as much IT capacity as it needs for a specific task.

Commercial cloud services promise to enable an organisation to scale-up and scale-down its IT infrastructure within minutes, rather than the weeks or months that would be required to commission or de-commission in-house infrastructure.

Why do commercial cloud providers have sufficient capacity available to do this? Primarily because they can amortise the cost of meeting peak demand across the globe and across different industry sectors. They can balance the peaks in one sector with troughs in another, meaning public cloud computers can achieve a higher level of utilisation that a standalone organisational IT infrastructure. For individual researchers, this translates into less time waiting in a queue for the necessary computing resources to become available, leading to potentially substantial productivity gains. In some cases, researchers relying on large shared facilities have to wait days, weeks or months to assess their data, slowing down the process of refining assumptions and the underlying model. The faster a researcher goes through the scientific method (see Figure 3), the more productive they are. In essence, cloud computing could remove the friction that can prevent some research from being completed in a timely fashion.

Figure 3: How cloud computing promises to speed up the scientific method (Source: AWS)

While such benefits have long been widely acknowledged in the research community, this kind of flexibility is becoming even more important as the volume of data being captured and analysed by scientists grows exponentially. It is becoming increasingly hard for dedicated in-house IT infrastructure to keep pace with demand.

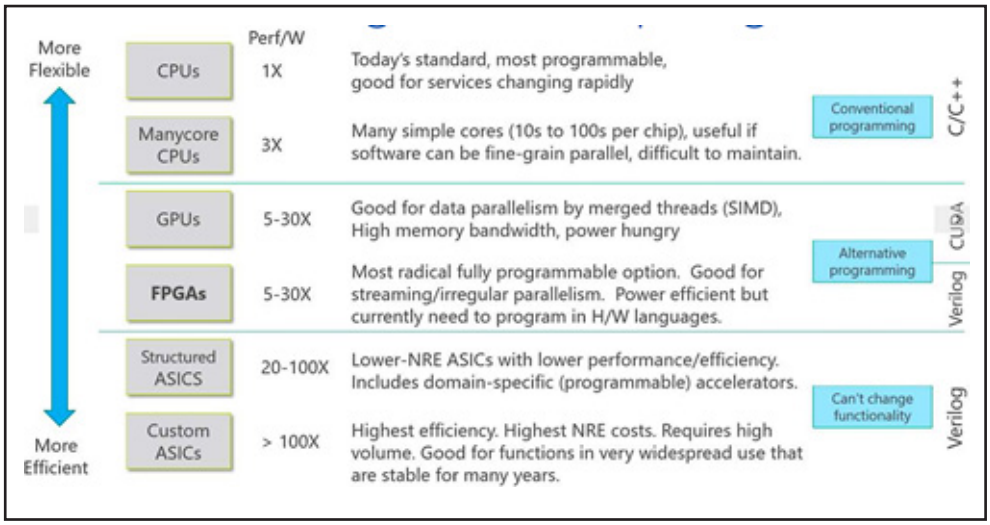

Figure 4: Different kinds of processors have different strengths and weaknesses (Source: Microsoft)

Cloud computing services can also be flexible in other ways. The architectural constraints of a legacy in-house IT infrastructure may mean researchers need to adapt their software and techniques to suit a platform configuration designed for traditional workloads. Moreover, the highly specialised nature of an in-house platform often means researchers can't modernise the software environment itself to allow for greater simplicity or better usability – all of which impacts their productivity.

By contrast, commercial cloud providers have diverse resources with very different characteristics designed to serve the needs of a diverse customer base. For example, different kinds of processors are suited to different kinds of workloads. Whereas standard CPUs are highly flexible, custom ASICs are highly efficient (see Figure 4). Easy access to a wide variety of computing resources creates scope for researchers to experiment on an ad-hoc basis with new techniques or tools. Cloud service providers say customers can develop entire end-to-end workflows with every component running on an optimal platform on an on-demand basis. On-demand optimisation should increase utility and reduce overall cost.

In a public cloud environment, complex workflows and complete environments can be stood up (or torn down) in minutes and shared widely, enabling rapid prototyping and development of new tools. This flexibility enables software to be developed rapidly.

For cash-strapped research institutions, financial flexibility can be as important as technical flexibility. Commercial cloud providers offer a range of different payment models. For example, the provider of the data could pay the cloud provider for the bandwidth required to enable other entities to access that data. In this case, the provider could cap the amount they are willing to pay each month. Alternatively, the provider could stipulate that the entity accessing the data will have to pay for the bandwidth required to access the data.

In recent years, public cloud service providers have adopted standardised hardware, creating further economies of scale as they can source servers and databases from a wide range of vendors. As more and more organisations migrate their IT into the public cloud, the major providers have created so-called hyper-scale clouds, with data centres that can be larger than dedicated HPC facilities and have far more servers.

The major cloud service providers now have the economies of scale to upgrade their data centres every few years, introducing new features and capabilities, and becoming more energy-efficient as time goes on. By contrast, legacy in-house IT architecture can be limited and energy-intensive.

Harnessing these continually evolving public cloud services should enable the EOSC to fulfil its goal of offering services at the "highest Technology Readiness Levels (TRLs)"

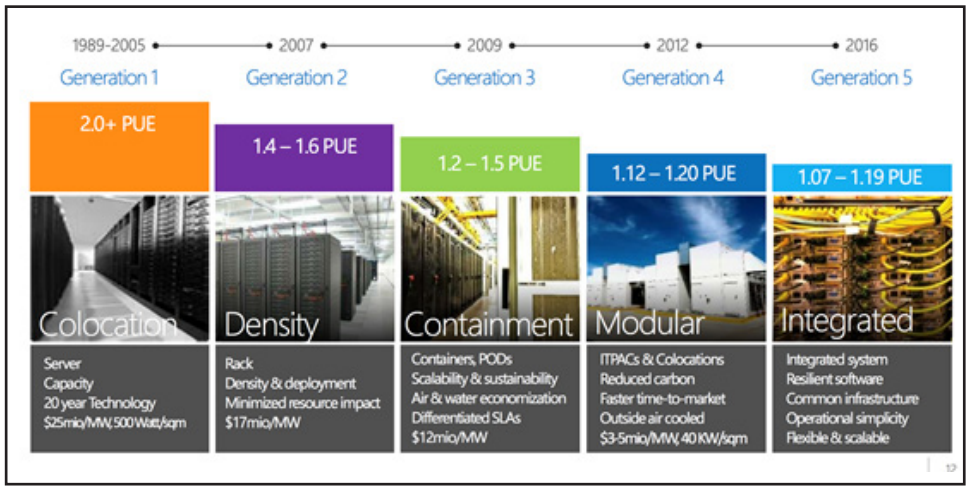

Figure 5: The rapid evolution in datacentres (source: Microsoft)

By providing a standardised service with a standardised contract in a multi-tenant architecture, hyper-scale cloud services can also bring economies of scale to compliance. A cloud service can be certified as compliant with global, regional, national and industry-wide standards, saving the end-user large sums in compliance costs. Otherwise, a research institution may have to spend a six-digit figure sum on one certification. This built-in compliance will increase the likelihood of the data set being exploited by the increasingly interconnected scientific community underpinned by the EOSC.

Efficiency gains arising from the EOSC

By making it much easier for scientists to access data generated by other research parties, the EOSC could dramatically improve productivity and efficiency, while increasing the overall value generated by each research project. Open access to data will allow scientists to contextualise and frame their research – they will be able to quickly identify the outstanding questions that need answering in their field. As a result, scientists will be far less likely to conduct overlapping and unnecessary research.

Moreover, with the EOSC, data can stay in one place – it won't need to move around. It can be located where it can be accessed by the computing power required to manipulate it and process it. Some experts argue that the cost of networking technology hasn't fallen as fast as cloud computing processing, meaning it is generally more cost-effective to move workloads, rather than move the data itself.

If the data remains in one place and is accessed via the cloud computing service's local network, the provider may not charge access costs or data egress fees, benefitting both the data provider and the data user. If the user doesn't need to create its own copy of the data, it won't need to buy storage for large datasets or pay for moving large volumes of data around.

The money used to acquire, store and maintain petabytes of data (hundreds of thousands, or millions of dollars), could be better spent on the processing time dedicated to unearthing an insight that leads to a discovery. For the public institutions funding research today, a big reduction in the number of data sets that are duplicated could amount to a major cost saving.

Moreover, users of the EOSC will be able to adapt and improve one another's workloads, increasing their efficiency and effectiveness, just as the open source community collectively improves software.

By making advanced computing resources available to all researchers, the EOSC could also reduce the growing gap between the scientific computer elite and the people emailing spread sheets.

By giving the latter group access to sophisticated tools, the EOSC could vastly increase the efficiency and effectiveness of the long tail of scientific research. Although many scientists can, in theory, access these resources, many lack the know-how, the tools and the techniques to employ very powerful hardware. The EOSC could change that by making such resources available through straightforward software tools.

In similar fashion, the European Open Science Cloud and the European Data Infrastructure could benefit businesses, including SMEs, which lack cost-effective and easy access to data storage, services and advanced computing.

One of the primary objections to the wider use of commercial cloud services for scientific research is the risk of vendor lock-in. Sceptics fear the financial and manpower costs of reformatting their data for a different cloud service will be so high that they won't ever be able to switch providers. These challenges are being addressed by a pre-commercial procurement of commercial cloud services for the research community, called Helix Nebula Science Cloud (HNSciCloud[1]), which is deploying a pilot hybrid cloud platform for end-users from a range of scientific disciplines.

The EOSC could further allay such fears by stipulating that public cloud service providers meet certain requirements and specifications that would enable an organisation to transfer research from one vendor to another in a cost-effective way. For example, the EOSC could employ mechanisms to avoid excessive concentration on one type or source of cloud resources and/or some kind of standard contract that ensures cloud providers meet service level agreements and observe portability standards.

1. http://www.hnscicloud.eu/

4. The broader benefits of the EOSC

In time, the EOSC should enable scientific breakthroughs that would not have been possible without open science. While it is impossible to put a financial value on these breakthroughs, scientific research has a long track record of delivering major economic benefits in the medium-to-long term. Think of the steam engine, electricity, antibiotics, oil refining or radio communications. In this context, anything that makes scientific research significantly easier could and should yield a return on the necessary investment for society as a whole. Whether that return will be 2X, 15X or 100X can't be predicted.

Conversely, scientific progress in Europe could slow without wider access to the data-crunching capabilities that the EOSC will provide. In future, researchers will become increasingly reliant on artificial intelligence and machine learning, which depends on vast amounts of data. As more and more science is conducted using artificial intelligence, researchers will need a combination of computational power, know-how and scientific data to drive innovations.

Individual European countries and regions don't have the necessary data or the computing resources to keep pace with the US and China in this regard. Datasets, in particular, need to be aggregated across Europe. One of the reasons China is becoming an increasingly important player in the development of artificial intelligence is the sheer volume of data available to its researchers and their neural networks.

The EOSC will also make it easier to track correlations in data from different disciplines. For example, European researchers are increasingly looking to combine genetic data, disease data, health records and socioeconomic data in health studies.Similarly, a combination of geographic information and human behaviour data has been utilised in service planning and construction both in the public and private sector. In many cases, services that draw data from multiple disciplines can become integral to every day life. For example, drivers now rely on real-time traffic maps (a combination of geographic and behavioural data) to get from one place to another as efficiently as possible.

The Commission believes the EOSC will also help researchers get their data skills recognised and rewarded. It should allow for easier replicability of results, while reducing rent-seeking. The initiative may also help address issues of data clearance and personal data protection.

In time, the mandate of the European Open Science Cloud and the European Data Infrastructure will be widened to serve the entire public sector, and ultimately industry. The EOSC should eventually ensure that all public data is fully discoverable, accessible and exploitable by scientists, policy makers and businesses. In this way, it could yield important new insights about how European society works and lead to the creation of thousands of innovative start-ups.

The development of Europe's entrepreneurial ecosystem has historically been held back by a cultural aversion to risk. The EOSC could help to address this issue by lowering the cost of failure: Collective action (such as pooling data resources and pay-as-you go computing) lowers the cost of accessing lots of data, hence lowering the risk.

In summary, the EOSC promises to have profound impact on European scientists' computing capabilities, giving them access to software tools, algorithms and data that they can't access today. As a result, scientists in both the public and private sectors will be able to conduct new kinds of experiments and research, with a lower level of risk, which could ultimately yield major economic benefits. The net effect would be to breathe new life into existing investments and draw new money into European science, creating a virtuous circle that fuels investment in innovative businesses and new public services.

5. Funding the EOSC

There are various sources of funds that could be used to meet the costs of establishing and maintaining the EOSC. As outlined in Chapter 2, the European Commission has allocated €260 million for the federation of the existing scientific data infrastructures. It sees this sum as the only additional funds that will be required beyond what is being spent today. To put such numbers in perspective: The Commission estimates that the EU (including the member states) spends approximately €10 billion a year on data infrastructures for science conducted in universities and other publicly owned facilities. That figure does not include HPC capacity.

More broadly, the European Commission estimates the public sector in Europe spends approximately €200 billion a year on scientific research, including the funding available through the EU framework programmes, such as Horizon 2020. If just 1% of that funding were to be allocated to data management, the overall data management budget that could be harnessed by the EOSC would rise by €2 billion per annum to €12 billion (assuming approximately €10 billion per annum is spent on data management today by publicly-owned research institutions). This sum could be used to meet the costs involved in the migration to the cloud, the opening up of data and enabling data interoperability, as outlined in Chapter 2.

If the EU and member states' research funding agencies insist that a small percentage of a research grant is used to pay for the professional storage, curation and management of data, funds that had been spent on ad-hoc data storage could be freed up for spending on compute cycles to analyse the data.

Rather than paying the entire research grant in cash, the funding agency could make a small percentage of it available as credits that can be spent on any kind of cloud service (so long as it meets the EOSC's technical/security/privacy criteria). This approach would help drive competition between cloud providers: the researcher's IT specialists would spend the credits with the provider offering the best value. Such a system would also allow for measurement of activity (since the EOSC/Commission would see where the credits were getting redeemed) and help ensure that European science isn't becoming overly dependent on a single cloud vendor.

Over time, the EOSC could also generate its own income stream by serving the needs of the private sector. There are many different ways in which the EOSC could be monetised, so the business model will need to be carefully conceived and refined over time. This will be the subject of a future report. However, in principle, the EOSC's funding mechanism should be designed to make public research as cost-effective as possible, while harnessing a share of the revenue generated by related commercial propositions.

Although the EOSC is intended to make research data free at the point of use for scientists, commercial entities could be required to pay to access data within the EOSC framework once their usage rises beyond a specific threshold. A points-based application process, designed to gauge the public value of the project in the broadest sense, could be used to determine the thresholds that apply to each entity. This approach may face opposition in those countries that consider the output of publicly funded research as a public asset that should be freely available to all.

But opponents may accept this kind of system if the thresholds were set high enough to ensure that they didn't act as a barrier to entrepreneurship and innovation.

For example, a university department engaged in fundamental research or a start-up with no revenues could be awarded a much higher level of free credit than the R&D unit of a large enterprise developing a commercial product. The free credit available to start-ups could taper off over time as they mature, so that the threshold is much higher in year one than in year five. Once they have exhausted their credits, a registered entity would then have to buy more credits from the EOSC's executive body. The funds raised in this way could be used to meet the running costs of the EOSC and further develop its capabilities and expand its scope.

Ultimately, the data within the EOSC could underpin an ecosystem of commercial services, just as the satellite data being captured by the European Space Agency is being used as the basis for commercial offerings (see next section).

Case study 1 – the European Space Agency

The European Space Agency has opened up the data generated by the Copernicus earth observation programme, one of the largest data providers in the world. Captured by the Sentinel satellites orbiting the earth, the data becomes freely available after an embargo in which the principal investigator has an exclusive window. When all the Sentinel satellites are operational, they will capture in excess of 10 petabytes of data each year.

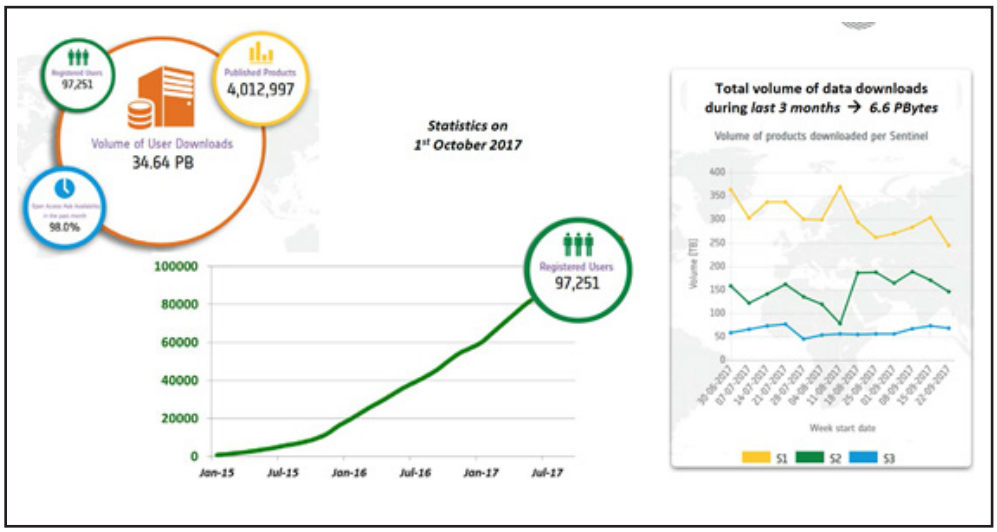

Figure 6: There is strong demand for data captured by the Sentinel satellites (Source: ESA)

Approximately 100,000 registered users have accessed data and images via the "SciHub" web portal, run by ESRIN, the ESA's Centre for Earth Observation (see Figure 6). Although Copernicus doesn't offer a data processing and analytics services to its users, commercial cloud services are emerging to fill the gap.

Indeed, the European Commission's new Copernicus DIAS (Data and Information Access Services) initiative is designed to kick-start the development of data access and cloud processing services, which can be used by entrepreneurs, developers and the general public to build Copernicus-based services and applications. The over-arching goal is to boost user uptake, stimulate innovation and the creation of new business models based on Earth observation data and information.

The plan is to enable users to access the Copernicus data and information close to processing facilities that can extract value from the data. To avoid duplication of data storage activities across Europe, the Commission is awarding service contracts that call on the cloud provider to offer free access to the Copernicus data, supplemented by cloud computing services on a pay-as-you-go basis.

To foster competition and the development of creative solutions, the Commission has awarded DIAS contracts to four consortia:

- Consortium led by Serco with OVH as cloud provider;
- Consortium led by Creotech Instruments with Cloudferro as cloud provider;
- Consortium led by Atos Integration with T-System International as cloud provider;
- Consortium led by Airbus Defence and Space with Orange as cloud provider.

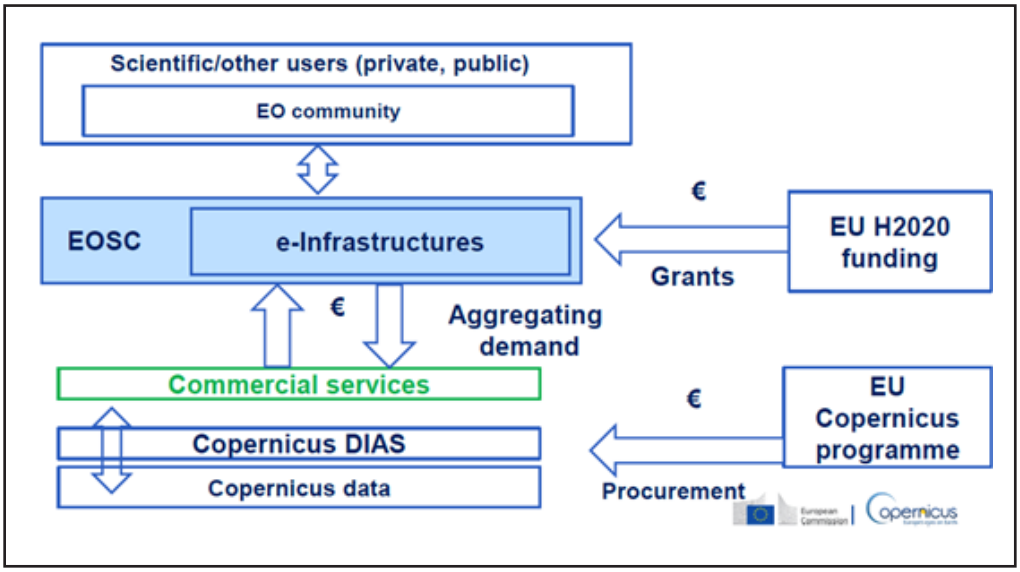

Figure 7: How the Copernicus DIAS programme could fit into the EOSC (Source: ESA)

EUMETSAT, in cooperation with ECMWF and Mercator Océan, are also implementing a DIAS, which means that by the second quarter of 2018, five DIAS will be available to users.

Competition amongst DIAS providers should stimulate innovation and avoid lock-in situations for the Commission and for users alike. Each DIAS will compose back-office infrastructure and interface services through which the user's front office components can connect to the back office infrastructure. As a scalable computing environment, the back-office will give unlimited, free and complete access to Copernicus data and information, and any other data that may be offered by the DIAS provider. However, the DIAS provider will be able to charge for computing and storage resources employed by the user. Moreover, the DIAS interface will be a set of commercial tools and services that can be employed by users to create their own applications. The Commission says "a significant number of consortia" have submitted their proposals for the DIAS.

In summary, users will have full and free access to Copernicus data and services through the DIAS, and will, at commercial conditions to be determined by the DIAS providers, be able to process the data and information to create services for their end users.

The ESA anticipates that the EOSC will enhance and extend the DIAS proposition by making it easier for researchers across Europe to access the Copernicus data. Indeed, if the EOSC was already up and running, DIAS services could simply be integrated into the open science cloud framework.

Case study 2 - the Cancer Genome Atlas and the International Cancer Genome Consortium

Two of the world's largest collections of cancer genome data are available at no cost to qualified researchers through Amazon Web Services' Public Data Sets program. Access to these petabyte-scale genomic data sets is expanding the research community and accelerating the pace of research and discovery in the development of new treatments for cancer patients, according to AWS. At the same time, the open availability of these data sets encourages researchers to make use of AWS analysis tools.

In 2015, AWS made the Cancer Genome Atlas (TCGA) corpus of raw and processed genomic, transcriptomic, and epigenomic data from thousands of cancer patients freely available to users of the Cancer Genomics Cloud, a cloud pilot programme funded by the National Cancer Institute in the U.S.

The International Cancer Genome Consortium (ICGC) PanCancer dataset generated by the Pancancer Analysis of Whole Genomes (PCAWG) study is also available on AWS, giving cancer researchers access to over 2,400 consistently analyzed genomes corresponding to more than 1,100 unique ICGC donors.

Access to TCGA and ICGC on AWS is administered by third parties, Seven Bridges Genomics and the Ontario Institute for Cancer Research, respectively. These partners have the rights to redistribute the data on behalf of the original data providers. The partners also curate and update the data over time. Once accepted, users are able to access the data via the CGC Web portal or use the CGC's API for programmatic access to the data.

As they no longer need to download and store their own copies of the data before they begin their experiments, researchers can work faster using a broader toolset hosted and shared by the community within AWS. Making the cancer genome data sets and tools available in the cloud is also enabling a greater level of collaboration across research groups, since they have a common place to access and share data. Amazon says researchers are also able to securely bring their own data and tools into AWS, and combine these with the existing public data for more robust analysis.

In the 15 months after the launch of the CGC, more than 1,900 researchers have registered on the platform, representing 150 institutions across 30 countries. In total, CGC users have deployed more than 5,000 tools or workflows and performed 80,000 executions, representing more than 97 years of total computation. There is significant collaboration among users, with an average of seven members per project on the platform.

6. Conclusions and Recommendations

Although they pale alongside the potential economic benefits, the financial costs of setting up and running the EOSC are significant. While existing budgets can be reallocated to cover most of the initial costs required to get the EOSC up and running, the science cloud may need to generate some revenues to enable it to invest in the development of the software tools, specifications and standards that will be required to enable the initiative to deliver on its potential. Given the value that the EOSC could bring to private sector research and product development, it should be able to eventually build up a substantial revenue stream.

By employing a system of credits with thresholds that can be honed over time, the EOSC could ensure that its services are free-at-the-point of use for academic researchers, while charging usage fees to businesses employing the EOSC to underpin commercial offerings. Of course, the EOSC could also be monetised in other ways: the business model, which will need to be carefully constructed, will be the subject of a future report.

However, another school of thought argues that the EOSC may not need to generate any revenues, as it will become self-sustaining in the same way that open source software is maintained by its community of users (typically with some support from large technology companies). In this scenario, individual researchers, empowered to employ whichever platform makes most sense to them, will then be doing nearly all their work using publicly developed and widely shared mobile workloads. As scientists re-use and enhance each other's workloads, they will be improving and expanding the EOSC, which will take on a life of its own in a similar way to the open source movement.

To ensure that the EOSC is both efficient and effective, it should seek to benefit from market dynamics and competition wherever possible. The EOSC can benefit from the ongoing competition between commercial cloud service providers, which has resulted in price reductions even as capabilities have improved. As much as possible, scientists should be able to use whichever cloud tool best serves their specific needs.

To maximise competition and flexibility, the EOSC should seek to harness all forms of cloud computing. It needs to be straightforward for both public institutions and private companies to provide researchers with services under the auspices of the open science cloud. The EOSC should ensure that researchers have all the information they need to make a fully informed choice about which cloud services and resources to use – transparency and simplicity is the key to a well-functioning marketplace. Transactions need to be simple and swift. Although research funders should insist that grantees make their data open and compatible with the EOSC, the grants should be agnostic about what cloud services they use to make their data findable, accessible, interoperable and reusable.

Moreover, to maximise the effectiveness of the money spent on the EOSC, investments in the initiative should be driven by demand, rather than a "build it and they will come" mentality. Demand is likely to be particularly strong for PaaS capabilities, which can help researchers develop the algorithms and software they need for their projects. As much as possible, the EOSC should not require data to be ported from one place to another – it is more efficient to store data in a single location and perform analytics in that location, rather than create multiple copies of a large data set.

SCIENCE|BUSINESS® NETWORK

Bringing together industry, research and policy

Industry

Amazon
Amgen
AstraZeneca
Dow Europe
Frontiers
GE
Google
Huawei
Microsoft
Nickel Institute
Novartis
Pfizer
Sanofi
Total
Toyota

Academia

Aalto University
Chalmers University of Technology
ESADE Business & Law School
ETH-Zurich
European University Association (EUA)
Karolinska Institutet
KTH Royal Institute of Technology
KU Leuven
Medical University of Warsaw
Norwegian University of Science and Technology (NTNU)
Politecnico di Milano
Sorbonne University
The Association of Commonwealth Universities (ACU)
Trinity College Dublin
TU Berlin
University College London
University of Amsterdam
University of Birmingham
University of Bologna
University of Eastern Finland
University of Luxembourg
University of Pisa
University of Twente
University of Warwick

Public organisations

Barcelona Supercomputing Center
Business Finland
CERN
Centre National de la Recherche Scientifique (CNRS)
The COST Association
Eureka
European Space Agency
Fraunhofer
Hospital Saint Joan de Deu
Innovate UK
Innovation Norway
Research Data Alliance Europe (RDA)
Republic of South Africa - Department for Science and Technology
Sant Joan de Deu - Barcelona Hospital
Trimbos Instituut

Consortia and EU projects

ACM Europe Policy Committee
ATTRACT
Deusto International Research School (DIRS)
ERC=Science²
GEANT
ICHOM - International Consortium for Health Outcomes Measurements