A Cloud on the 2020 Horizon

Commission High Level Expert Group on the European Open Science Cloud

Realising the European Open Science Cloud: first report and recommendations

Preface by Barend Mons, Chair

This report aims to lay out a high level, living roadmap for the realisation of the European Open Science Cloud (EOSC). The High Level Expert Group, with ten members from European countries, Japan and Australia, discussed extensively in several meetings, conferences, policy events and met with key stakeholders (30 November 2015) and research funders (15 March 2016). Based on these consultations, on many 'white papers' and on a range of presentations and feed-back at international meetings, we are confident that our recommendations count on a high-level of consensus amongst all stakeholders. This was a solid basis to embark on this challenging journey with the Commission, the Member States and International partners in concert.

The title of this first report may have a slightly threatening ring to it and indeed, if we do not act, there might be a looming crisis on the Horizon. The vast majority of all data in the world (in fact up to 90%) has been generated in the last two years. Computers have long surpassed individuals in their ability to perform pattern recognition over large data sets. Scientific data is in dire need of openness, better handling, careful management, machine actionability and sheer re-use. One of the sobering conclusions of our consultations was that research infrastructure and communication appear to be stuck in the 20th century paradigm of data scarcity. We should see this step-change in science as an enormous opportunity and not as a threat. The EOSC is a positive 'Cloud on the Horizon' to be realised by 2020. Ultimately, actionable knowledge and translation of its benefits to society will be handled by humans in the 'machine era' for decades to come, machines are just made to serve us.

But let's not ignore the facts: the science system is in landslide transition from data-sparse to data-saturated. Meanwhile, scholarly communication, data management methodologies, reward systems and training curricula do not adapt quickly enough if at all to this revolution. Researchers, funders and publishers (I always thought that meant making things public) keep each other hostage in a deadly embrace by continuing to conduct, publish, fund and judge science in the same way as in the past century.

So far, no-one seems to be able to break this deadlock. Open Access articles are indispensable but solve only a fraction of the problem. Neither 'open research data' alone will do. We still try to press petabytes of results in length-restricted narrative, effectively burying them behind firewalls or in 'supplementary data behind decaying hyperlinks and then trying to mine them back again. Computers hate ambiguous human language and love structured, machine actionable data, while machine readable data are a turnoff for the human mind. As computers have become indispensable research assistants, we better make what we publish understandable to them. We need both in concert to form social machines; in order to do pattern recognition in complex, interlinked data as well as confirmational studies on methodology and rhetorics in plain understandable human language.

We hope that this report will be part of a game-changing effort of all European Member States and our international partners towards true Open Science.

It has been an enormous pleasure to work with the members of the HLEG, with great support from Commission staff; I also wish to acknowledge the continuous and open discussions and advice from colleagues from the United States.

Note added in press [date]: The European Commission launched the European Open Science Cloud as integral and main part of the Communication: European Cloud initiative, on 19 April 1026[1]. The Communication advances a number of specific actions needed to move from vision to action on the European Open Science Cloud by 2020. This report was drafted as a live, interactive think-box for the Commission in this drafting process; in that respect, we believe that it has accomplished its mission. Upon its publication today, the report provides a high-level roadmap for the successful realisation of the EOSC as part of a global research data commons.

[1] http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266

Members of the Commission High Level Expert Group on the European Open Science Cloud

Paul Ayris

Jean-Yves Berthou

Rachel Bruce (Rapporteur)

Stefanie Lindstaedt

Anna Monreale

Barend Mons (Chair)

Yasuhiro Murayama (Observer, Japan)

Caj Södergård

Klaus Tochtermann

Ross Wilkinson (Observer, Australia)

Executive Summary

The European Open Science Cloud (EOSC) aims to accelerate and support the current transition to more effective Open Science[2] and Open Innovation[3] in the Digital Single Market[4]. It should enable trusted access to services, systems and the re-use of shared scientific data across disciplinary, social and geographical borders. The term cloud is understood by the High level Expert group (HLEG) as a metaphor to help convey both seamlessness and the idea of a commons based on scientific data. This report approaches the EOSC as a federated environment for scientific data sharing and re-use, based on existing and emerging elements in the Member States, with light-weight international guidance and governance and a large degree of freedom regarding practical implementation. The EOSC is indeed a European infrastructure, but it should be globally interoperable and accessible. It includes the required human expertise, resources, standards, best practices as well as the underpinning technical infrastructures. An important aspect of the EOSC is systematic and professional data management and long-term stewardship of scientific data assets and services in Europe and globally. However, data stewardship is not a goal in itself and the final realm of the EOSC is the frontier science and innovation in Europe.

Challenges and general observations

The majority of the challenges to reach a functional EOSC are social rather than technical.
The major technical challenge is the complexity of the data and analytics procedures across disciplines rather than the size of the data *per se*.
There is an alarming shortage of data experts both globally and in the European Union.
This is partly based on an archaic reward and funding system for science and innovation, sustaining the article culture and preventing effective data publishing and re-use.
The lack of core intermediary expertise has created a chasm between e-infrastructure providers and scientific domain specialists.
Despite the success of the European Strategy Forum on Research Infrastructures (ESFRI), fragmentation across domains still produces repetitive and isolated solutions.
The short and dispersed funding cycles of core research and e-infrastructures are not fit for the purpose of regulating and making effective use of global scientific data.
Ever larger distributed data sets are increasingly immobile (e.g. for sheer size and privacy reasons) and centralised HPC alone is insufficient to support critically federated and distributed meta-analysis and learning.
Notwithstanding the challenges, the components needed to create a first generation EOSC are largely there but they are lost in fragmentation and spread over 28 Member States and across different communities.
There is no dedicated and mandated effort or instrument to coordinate EOSC-type activities across Member States.

Key factors for the effective development of the EOSC as part of Open Science

New modes of scholarly communication (with emphasis on machine actionability) need to be implemented.
Modern reward and recognition practices need to support data sharing and re-use.
Core data experts need to be trained and their career perspective significantly improved.
Innovative, fit for purpose funding schemes are needed to support sustainable underpinning infrastructures and core resources.

[2] For background and policy context, see http://ec.europa.eu/research/openscience/index.cfm
[3] Carlos Moedas speed, http://europa.eu/rapid/press-release_SPEECH-15-5243_en.htm
[4] http://ec.europa.eu/priorities/digital-single-market/

A real stimulus of multi-disciplinary collaboration requires specific measures in terms of review, funding and infrastructure.
The transition from scientific insights towards innovation needs a dedicated support policy.
The EOSC needs to be developed as a data infrastructure commons, that is an eco-system of infrastructures.
Where possible, the EOSC should enable automation of data processing and thus machine actionability is key.
Lightweight but internationally effective guiding governance should be developed.
Key performance indicators should be developed for the EOSC.

Specific recommendations to the Commission for a Preparatory Phase

Policy recommendations

P1: Take immediate, affirmative action on the EOSC in close concert with Member States
P2: Close discussions about the 'perceived need'
P3: Build on existing capacity and expertise where possible
P4: Frame the EOSC as the EU contribution to a Internet of FAIR Data and Services underpinned with open protocols

Governance recommendations

G1: Aim at the lightest possible, internationally effective governance
G2: Guidance only where guidance is due (this relates to technical issues, best practices and social change).
G3: Define Rules of Engagement for service provision in the EOSC
G4: Federate the gems and amplify good practice

Implementation recommendations

I1: Turn the HLEG report into an EC High level Roadmap to scope and guide the EOSC initiative
I2: Develop, endorse and implement the Rules of Engagement for the EOSC
I2.1: Set initial guiding principles to kick-start the initiative as quickly as possible
I3: Fund a concerted effort to develop core data expertise in Europe
I4: Develop a concrete plan for the architecture of data interoperability of the EOSC
I5: Install an innovative guided funding scheme for the preparatory phase
I6: Make adequate data stewardship mandatory for all research proposals
I7: Install an executive team to deal with international coherence of the EOSC
I8: Establish an executive team to deal with the early preparatory phase of the EOSC

The European Open Science Cloud? Some nuances and definitions

Imagine a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes. Imagine that this all operates under well-defined and trusted conditions, supported by a sustainable and just value for money model. This is the environment that must be fostered in Europe and beyond to ensure that European research and innovation contributes in full to knowledge creation, meet global challenges and fuel economic prosperity in Europe. This we believe encapsulates the concept of the *European Open Science Cloud (EOSC)*, and indeed such a federated European endeavour might be expressed as the European contribution to a Internet of FAIR Data and services.

The European Open Science Cloud is a *supporting environment for Open Science* and not an '*open Cloud*' for science.

The EOSC aims to accelerate the transition to more effective Open Science and Open Innovation in a Digital Single Market by removing the technical, legislative and human barriers to the re-use of research data and tools, and by supporting access to services, systems and the flow of data across disciplinary, social and geographical borders. The term European Open Science Cloud requires some reflection to dispel incorrect associations and clarify boundaries; in fact the term 'cloud' is a metaphor to help convey the idea of seamlessness and a commons.

- **European**: research and innovation are global. The EOSC cannot be built exclusively in and for Europe. Serious efforts are needed to ensure coordinated action with other regions. Europe, being inherently federated, is in a strong position to lead this initiative.

- **Open**: the use of Open in relation to research has been widely discussed over recent years, and it is acknowledged that not all data and tools can be open. There are exceptions to openness, such as confidentially and privacy. Open is also often confused with 'for free'. Free data and services do not exist[5]. These nuances need to be respected and intelligently open is what we mean, often referring more to accessibility under proper and well defined conditions for all elements of the EOSC[6].

- **Science**: the use of the term science explicitly includes the arts and humanities, and in fact no current or future discipline should be excluded from the EOSC. In addition the Science Cloud infrastructure should support not only innovative scientific research but also societal innovation and productivity, which takes place predominantly in collaboration between research institutes and the private sector. The EOSC should also support broad societal participation in Open Innovation and Open Science.

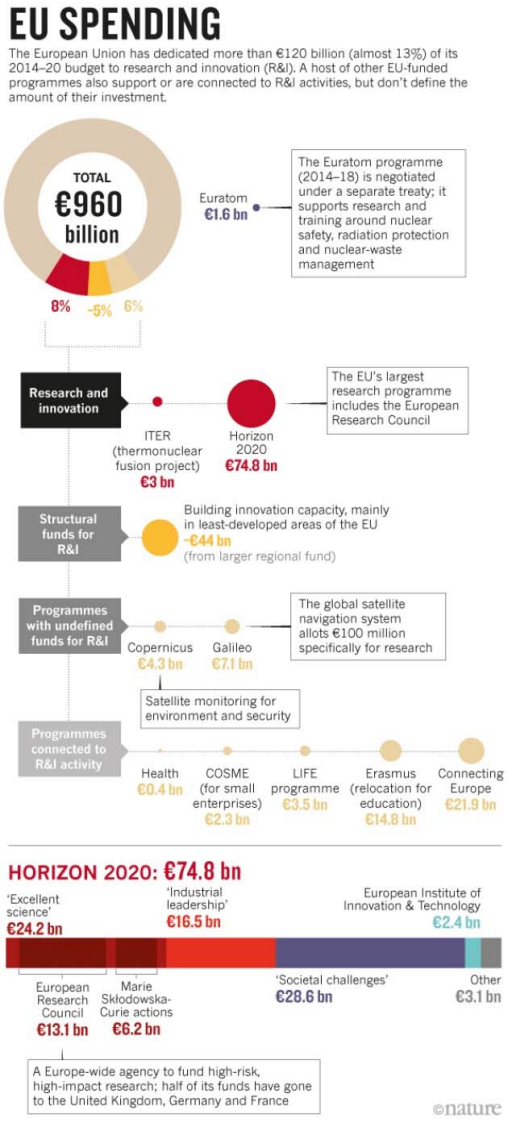
- **Cloud:** the term cloud can cause considerable confusion as it has many connotations. It can be misinterpreted to indicate that the EOSC is mostly about hard ICT infrastructure and much less about a commons of data, software, standards, expertise and policy related to data-driven science and innovation.

[5] Although scientists may perceive scientific data services that are at no cost to them to be free, and oppose commercial approaches even if they are demonstrably better then free alternatives.

[6] See for basic principles the UK report Science as an Open Enterprise, https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/

The European Open Science Cloud in the context of Open Science

The EOSC is a need emerging from science in transition[7]**.** The EOSC is indeed European, but it should also be interoperable with the Internet of FAIR data and services and be an accessible infrastructure for modern research and innovation. It includes the required human expertise, resources, standards, best practices and underpinning infrastructures. It will have to support the Finding, Access, Interoperation and in particular the Re-use of open, as well as sensitive and properly secured data. It will also have to support the data related elements (software, standards, protocols, workflows) that enable re-use and data driven knowledge discovery and innovation. An important aspect of the EOSC is therefore professional data management and long term data stewardship. The latter aspect is presently lacking.



Mostly due to current methods capture and data malpractice, approximately 50% of all research data and experiments is considered not reproducible, and the vast majority (likely over 80%) of data never makes it to a trusted and sustainable repository. At an investment of Europe in data-generating research of €120B between 2014-2020,[8] the annual capital destruction is consequently very substantial[9]. This does not take into consideration associated losses from inefficient data analysis and the economic impact of stalling innovation and societal non-applicability of knowledge. Good data stewardship and a globally operational Internet of FAIR data and services will significantly reduce these losses and fuel science and innovation.

Europe enjoys a long tradition and a relatively healthy research infrastructure, served via domain specific European Research Infrastructures and cross-domain ICT e-infrastructures, as well as other disciplinary and cross disciplinary collaborations and services. Many Member States also provide infrastructures and initiatives that support research and data access and use. Although these were largely built in the earlier phases of the data revolution, they are nevertheless important foundations for the EOSC and should be built upon.

However, a step change is required to realise the ambition of increased seamless access, reliable re-use of data and in fact all digital research objects[10] and collaboration across different services and infrastructures, where data access and re-use is open to all actors across public and private spheres. This will mean a new way of working through deep, equal partnerships between the science communities and the ICT communities so that the EOSC can optimally benefit from all expertise.

[7] The following points are all supported by a range of recent policy and position papers by stakeholders. These will be placed online alongside this report.

[8] Nature [to do]

[9] a: 90% of world's data generated over last two years (http://www.sciencedaily.com/releases/2013/05/130522085217.htm)
b: US$28B/year (50%) spent on preclinical research is not reproducible: Freedman et al. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165
c: Only 12% of NIH funded datasets are demonstrably deposited in recognised repositories: Read et al. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132735

[10] See Glossary.

Science itself is in an unprecedented phase of transition, driven by the power of rich and complex data, networked digital technology and its ability to underpin new approaches to research, knowledge management and innovation. As a consequence, practices, social structures and infrastructures that have gradually developed over centuries, now need to undergo a step-transition. Many of these practices are rooted deeply in the scientific community and in the support structures of research and they appear to be quite resilient to this required change. A fundamental shift nevertheless is needed to match the potential to generate ever increasing amounts of data and to turn these data into knowledge as renewable, sustainable fuel for innovation in turn to meet global challenges. This transition was marked by the EC as a transition to Open Science[11]. The EOSC is an environment that needs to be realised to underpin and enable this transition.

Key trends of Open Science and their relevance for the EOSC

New modes of scholarly communication: scholarly communication, which has been dominated by narrative and verbal means of delivery for centuries, should be moving rapidly towards communication and re-use formats that also better suit our main research assistants: the data generating machines and data processing machines.

Modern rewards and recognition: assessment, selection, funding and reward systems in research have to be urgently adapted and updated. The current systems, mainly based on the data-sparse and narrative ages, strongly bias the science system towards narrative publishing and new -publishable- tool generating research. The current system provides little if any support and incentive for data publishing and for tool sharing nor for the development and reward of data related expertise; data stewardship and (re-) analysis to support the final aim of science: knowledge discovery.

Increasing reliance on data experts, especially in academia where they are most severely undervalued, a lack of data related core expertise may well be among the risks for Europe losing a leading position in science. New forms of output publication for data and software are emerging that need to be given credit in research assessment and as part of promotion decisions if we are to support the change towards open and data-driven science.

Cross-disciplinary collaboration: cross-disciplinary collaboration is critically needed, as scientists increasingly use raw and curated data resources and analytics tools from disciplines other than their own. However, currently, other people's data is notoriously difficult to discover even within one's own discipline. Discovering relevant (other people's) data from other disciplines will be even more difficult. For example health researchers now want to use data from social media and the 'quantified self'[12]. But, how would a health researcher know about a valuable dataset in say, the humanities, when terminology, data formats and meta-data standards are completely different? With the current absence of proper meta-data standards and the lack of data and tool search engines researchers cannot be blamed for re-inventing a new wheel. This is further amplified across disciplines. Use of text and data mining techniques are essential for the EOSC, research analysis and support of cross-disciplinary use. However too often there are legal barriers. In Europe this is currently the subject of discussion in the Commission and it is likely we will see a pan-European exception for text and data mining in 2016.

Fostering transition from science to innovation: Although severely sub-optimal, knowledge discovery nevertheless has reached such a pace that the translational and innovation capacity of society has difficulty to keep pace. Especially in Europe, where the support for one of the most innovative elements in society; SMEs, is relatively weak. Multidisciplinary research and innovation projects and

[11] ec.europa.eu/research/openscience/index.cfm; other terms considered include: as Science 2.0', data driven science', participatory science', 'science highway', 'better science', 'open research' and 'open scholarship' – the latter two were included as alternatives to the word 'science', which could be interpreted as excluding the humanities in some cultural contexts.
[12] https://en.wikipedia.org/wiki/Quantified_Self

public-private consortia are supported on paper in more policy-papers than we can possibly read, but in *actual practice* the European financing and review climate is severely hampering the actual flourishing of these crucial partnerships.

A complex eco-system of infrastructures: it may seem counter-intuitive but the challenges of ever bigger data can no longer be solved only by ever bigger infrastructure. Next to advanced computer science, which will bring innovative computing and storage and new advanced algorithms for knowledge extraction from data, we need *fundamentally* to rethink infrastructure as we know it. With the growth of data in more and more disciplines outpacing the increase of transfer speed, many comprehensive datasets are simply too big to move efficiently from one location to another. Moreover, data are in many cases so privacy sensitive that legislation effectively precludes their moving outside the environment in which they have been collected[13]. Therefore, relatively lightweight workflows (e.g. process virtual machines) containing parallel and distributed analytics algorithms increasingly visit data where they reside, with supporting reference data and transporting only conclusions outside the safe data vault. This is an early instantiation of an internet of data and services where containers with software applications are routed to relevant data and vice versa. This approach will unleash enormous distributed analytics power, but there are intellectual challenges to address and the hardware containing the data must have tailored and appropriate high throughput compute (HTC) capacity connected to them. Centralised supercomputing locations that are crucial for solving high capacity HPC scientific challenges alone will not adequately support this irreversible trend. Complementary infrastructures are needed.

Machine understanding: the size and complexity of many data sets is such that only powerful computers can process them and reveal patterns that may lead to actionable knowledge extraction by and for human users. Machines have become essential research assistants, both for data generation, data processing and analytics. Data formatting, terminology/identifier mappings and provenance must therefore be optimally organised in order to support machine processing as well as the human-mind knowledge extraction. However, the tools supporting these two processes are fundamentally different, pattern recognition tools being mainly for machines and tools for confirmational reading and interpretation being mainly for humans. Machine actionability of whatever is published[14] is therefore a crucial consideration in modern narrative and data publishing.

Research integrity: there is an alarming lack of reproducibility of current published research, together with scientific fraud, this causes enormous damage to the reputation of science. This is partly due to the lack of deep and rigorous knowledge on how to render data and the associated methodology and tools in a format that allows others to reproduce results. An important nuance is that not all conclusions in the literature for which the results cannot be easily reproduced elsewhere are wrong, but reproducibility and early detection of fraud-signals and optimal re-use of assets will increase as a result of good data stewardship and core data processing and analysis expertise.

[13] Data in the 100,000 Genomes Project (version 1.0 02/09/15) https://www.youtube.com/watch?v=nneWFaJ6Hfc

[14] SDATA-15-00190: Wilkinson et al 2016: Wilkinson et al: FAIR Data: Guiding Principles for Scientific Data Management and Stewardship: http://www.nature.com/articles/sdata201618

A clash of cultures?

A 'clash of cultures' emerged as a key challenge[15] from the stakeholder consultation, as historical developments may have led to a chasm between domain specialists and e-infrastructure specialists. Open Science drives two very different communities and cultures together and we lack connective tissue. In the earlier transition of scientific research from a largely individual, elite and intellectual activity to a mainstream, institute-based activity, a new profession emerged: the research analyst who soon became indispensable and highly recognised research professionals, in academia as well as industry. They were co-publishing and involved in all aspects of the research and throughout the experimental cycle. When data-generating machines became mainstream and high-throughput data generation boomed, the experts who knew how to operate these data *generating* machines only gained in importance. However, computer and data specialists – new key actors in modern research – who knew how to operate data *analytics* machines were not given the same credit in the scientific research process as wet-lab analysts in the past.

These new data experts come from scientific and engineering cultures with very different reward systems and incentives[16], different jargons and very different skill sets. These cultural differences resulted in understandable but unnecessary mismatches and in an alarming scarcity and loss of crucial data-related skills in research. This, in turn, has created a divide between researchers and those that support research with data processing and software. As a consequence, these two communities that are both essential to Open Science have not closely co-evolved. Often, frontline ICT development for science takes place rather independently from day to day science practice. In contrast to other lab-equipment or research infrastructure, scientists frequently misjudge ICT infrastructures as supporting infrastructure, a commodity or tool that can be relatively easily purpose-built; they underestimate the complexity and the need for professionalism. The added value of working together is not obvious and it deserves more recognition.

In addition, financial support for data generating scientific activity and support for the underpinning research infrastructure have traditionally been separated, both in many Member States and at the EC level. This may have aggravated a separation of worlds, a major challenge emerging from the consultation process, as it has undermined communication and collaboration between the top-tier ICT experts and top-tier experimental and other scientists. The few examples where these professionals have been able to bridge the divide and have effectively collaborated are positive exceptions to the rule. Most researchers are still struggling forward with severely suboptimal solutions, sometimes out of ignorance of what is available, but often because actual collaboration with computer scientists and engineers takes time and is not easy.

This complex state of affairs results in fundamental misunderstandings between the current main stakeholders in ICT and e-infrastructure providers and scientific domain-users. The complexity, cost and intellectual challenges in the other domain are systematically underestimated and undervalued in both directions[17]. This hampers the user-driven and expert co-development of core scientific data infrastructures, which is required so that Open Science can meet its full potential.

Moreover, there are very few incentives in the reward system for user support and for the support for data or tool re-use, such as proper documentation of code, versioning and scalability considerations. Scientists expect that Open Source, project-funded software tools will stay magically updated, online; they will continue to produce and consume new data types, even when the project generating them is long-gone. Furthermore, once ICT infrastructures and tools become commodities many scientists do not see the logic of co-authorship of the data analysts on scientific publications and only acknowledge them for analysing the data.

Finally, the divide frequently precludes agile co-development with continuous power-user feedback and rigorous testing of prototypes is also precluded as domain-users do not always conceive of the possibilities the latest developments in ICT enable. This is contributing to the lack of data scientists that venture out from classical computer or data science departments into other scientific fields.

[15] See online additional material at http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

[16] For example the publication for IEEE conferences versus first or last author on papers in high impact journals (which is not the same as high impact papers, let alone high impact research).

[17] This is known as the Dunning-Kruger effect, https://nl.wikipedia.org/wiki/Dunning-krugereffect

Data expertise is lacking in the EU

As a side effect of the above, there is an alarming shortage of data expertise in the EU and a pressing requirement with regards to the data expertise needed to support the aims of the EOSC is apparent. It became clear-and has been reflected in nearly all stakeholder contributions to the HLEG- that there is a major hole in the EOSC planning if we do not repair the significant lack of Core Data Experts. We use the term Core Data Experts here deliberately, emphasising that we are dealing with a range of skills that warrant the definition of a *new class of colleagues* with core scientific professional competencies and the communication skills to fill the gap between the two cultures.

Core data experts are neither computer savvy research scientists - although the latter also need to be educated to the point where they hire, support and respect Core Data Experts - nor are they hard-core data or computer scientists or software engineers. They should be technical data experts, though proficient enough in the content domain where they work to be routinely consulted in the research team at the very beginning (experimental design, proposal writing) until the very end of the data discovery cycle. They will work to secure that good data management plans are an essential part of good research practice (including data re-use and stewardship planning and proper budgeting) and the proper capturing of new data capture (formats, metadata richness, standards, provenance, publishing, linking and analysis), they will also support analysis. This package of skills and expertise is rare and the few people with this skill set are often attracted to industry or outside Europe where they are more respected and valued.

The number of people with these skills needed to effectively operate the EOSC is, we estimate, likely exceeding half a million within a decade. As we further argue below, we believe that the implementation of the EOSC needs to include instruments to help train, retain and recognise this expertise, in order to support the 1.7 million scientists and over 70 million people working in innovation[18]. The success of the EOSC depends upon it.

How will the European Open Science Cloud be realised?

Policy, Governance, first phase implementation and guiding principles

If *one* strong consensus arose from the consultations, policy papers and debates with the stakeholders it is that the EOSC should *not* be a new major, localised and centrally governed initiative. We believe that discussions and broad agreements on minimal standards and early rules of engagement were by default a first step in the realisation of the EOSC to be technically conceived as an Internet of FAIR data and Services.

We notice a clear analogy here with the early days of the Internet. The creation of NSFNET, choice of the TCP/IP standard and the authorised development of Domain Names enabled the boom of the Internet in the 1990's, where the development of the HTTP and HTML drove its major application domain, the largely textual WWW. This combination of authorisation, key support by a major leading agency (NSF) and a dedicated community (W3C) setting and enforcing minimal standards allowed virtually everyone to start building standard-compliant tools and services in the ecosystem. The Internet still has no major centralised governance in either technological implementation or policies for access and usage; each constituent network sets its own policies. Still, the early shaping of it and the openness of standards effectively prevented a situation where a few privately owned companies or public parties could entirely dominate and monopolise the developing Internet[19]. Internet standards were so minimal and rigorous that even after 25 years of overwhelming growth and development, they are still essentially the same. Only recently was there a move to IPV6 to allow a much broader range of IP addresses, which does not fundamentally change

[18] http://ec.europa.eu/eurostat/statistics-explained/index.php/R_%26_D_personnel

[19] Hart, Strawn, A Brief History of NSF and the Internet, August 2003, https://www.nsf.gov/od/lpa/news/03/fsnsf_internet.htm (and -paid book-) https://mitpress.mit.edu/books/inventing-Internet

the structure. When XML and RDF developed in a first attempt to develop schema free and self-defining components in the internet, nothing fundamentally changed.

We will need to stay as closely to the lessons learned and the choices made for the hypertext applications on the Internet and the early stage semantic web as well as the early developments in the Internet of Things. The EOSC will succeed only if it will have low barriers to data and services interoperability and a light-weight governance and financing structure.

At the European policy- and organisational level the EOSC should take an approach similar to that of the successful ESFRI roadmap where a preparatory phase is followed by an implementation phase. However, to meet the step change and ambition of the EOSC a more agile approach is required and so there are some key differences with the ESFRI approach. For instance we cannot afford a preparatory phase of many years, as the need of many disciplines for an early functional EOSC is very clear. It emerged from all position papers that most elements have been judged as being there but hidden in fragmentation. The consultation also highlighted many good practices, which the EOSC should work to amplify and build upon.

We need to commence defragmentation actions immediately, including the setting up of light and appropriate guidance and governance structures and prototyping as well as implementation for new solutions that are needed during the preparatory phase. We recognise that paths to implementation are not yet crystal clear and we heard from various stakeholders that in some cases the tough economic climate will need to be considered as well as the diverse practices, policies and funding for data stewardship exist across Member States and disciplines. A lightweight, agile and phased approach will be critical for rapid, tangible progress. The recommendations that follow are mainly for this proposed preparatory phase.

Recommendations of the High Level Expert Group

Policy recommendations

P1: Take immediate, affirmative action on the EOSC in close concert with Member States

Our first and overarching recommendation is that in order for Europe to have a modern and thriving research and innovation environment it is essential that Member States, internationally collaborating through the current instruments take immediate and solidly supported affirmative action to realise the first phase of a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes under well defined, secure and trusted conditions, supported by a sustainable and just and value-for-money model.

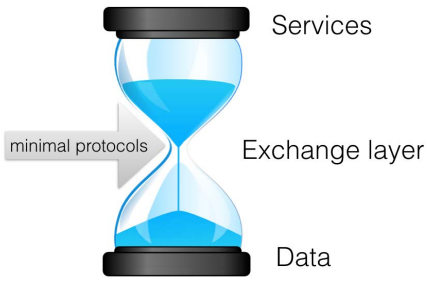
P2: Close discussions about the 'perceived need'

Although the EOSC ultimately will need to involve all ESFRIs, all e-infrastructures, private industry and stakeholders from Member States and beyond, the preparatory phase needs not be long. It is true that often preparation is required for landscape inventories and consensus building, as was the case for most of the ESFRIs. However in the case of the EOSC there has been a long consultation phase and we have no lack of position papers from stakeholders (see online accompanying material to this report), and there is no lack of consensus on the extremely urgent need for it.

P3: Build on existing capacity and expertise where possible

It is clear from the position papers and other inputs that for an overwhelming majority the elements of the EOSC exist and are of high quality; the issue is that they are hidden in fragmentation. Of course this does not mean that there are no major challenges left, but these are mainly cultural and reasonably well defined. As well as about infrastructure and services for open science, we learnt about many examples of good practice in terms of collaboration and policy that can be built upon. We therefore believe that many of the actions we propose for the preparatory phase of the EOSC can be significantly progressed or even completed by the end of 2017.

P4: Frame the EOSC as the EU contribution to a Internet of FAIR Data and Services underpinned with open protocols



The current internet application domain has avoided the dominance of a very limited number of private or public parties. Its 'hourglass model' with minimal, rigorous standards and protocols and maximum freedom of implementation has major advantages. We strongly advise to follow a similar approach to implement the EOSC which is based on minimal rigorous standards. It will allow open and common implementation and so it will prevent costly and time consuming exercises to decide who has the best solutions. Instead, it will allow participation from all stakeholders, including research infrastructure providers, Member States, research institutes and businesses. All providers, public and private, can start implementing prototype applications for the Internet of FAIR data and Services on the day minimal standards and the minimal rules of engagement are released.

Governance recommendations

G1: Aim at the lightest possible, internationally effective governance

Given the urgency and the number and variety of stakeholders and participants required to realise the EOSC, a tightly governed, new infrastructure built 'somewhere' is not the right model for the EOSC to be a success. Instead a more inclusive, flexible, transparent and less centralised approach is required, one that also enables effective global collaboration. The Commission needs to establish a lightweight, sustainable and collaborative governance model for the EOSC for all players to contribute.

G2: Guidance only where guidance is due

While we advocate lightweight governance we need a degree of regulation. For instance the harmonisation of the current 'standards jungle' needs to be actively coordinated. With no regulation, some major players, public and private, may claim an unjust and counterproductive share in the EOSC. The EOSC will have a myriad of small and very large players, as is the case in the current internet, but it should be perceived by regulators and stockholders alike as a commons where citizens, researchers and innovators need to use each other's data and tools in a trusted affordable and sustainable environment. Europe should take a lead in this due guidance element of the Internet of FAIR Data and Services.

G3: Define Rules of Engagement for service provision in the EOSC

To support wide participation, innovation and sustainability the EOSC needs to be open to all players, public and private, European and non-European and the development of the desired expert infrastructure will be guided and governed by a minimal set of rigorously applied and enforced protocols and developed by parties that endorse so called Rules of Engagement (RoE) that specify the conditions under which stakeholders participate. These RoE can be used to brand providers in the EOSC as trustworthy and compliant with the RoE, comparable to Conformant Cloud Providers in the USA[20]. It should be clear that non-EOSC approved players are free to explore any role in the Open Science ecosystem they wish, even if they do not adhere to the RoE. They will just not be able to brand their services as EOSC approved/certified[21].

G4: Federate the gems (and amplify good practice)

Based on the consensus that most foundational building blocks of the Internet of FAIR data and Services are operational somewhere, but that they operate in silos per domain, geographical region and funding scheme, we recommend that early and strong action is taken to federate these gems. Optimal engagement is required of the e-infrastructure communities, the ESFRI communities and other disciplinary groups and institutes. Several of these cross-ESFRI building blocks begin to operate in individual Member States. Simultaneously, the wealth of small and large industrial players in Europe should be engaged. All partners and stakeholders that adhere to standards and sign off on the RoE should be eligible.

Implementation recommendations

I1: Turn the HLEG report into an EC High level Roadmap to scope and guide the EOSC initiative

This report can be the basis for a formal paper to be approved by the Commission and subsequently endorsed by Member States. The resulting report can serve as a high level guiding document for actual developments and implementations in the Member States and Horizon 2020, as well as a discussion basis for further international consensus building and collaboration.

[20] https://grants.nih.gov/grants/guide/notice-files/NOT-LM-16-002.html

[21] Comparable to https://en.wikipedia.org/wiki/CE_marking
CE marking is the manufacturer's declaration that the product meets the requirements of the applicable EC directives.

I2: Develop, endorse and implement the Rules of Engagement for the EOSC

The Commission, in close collaboration with appropriate stakeholders in Member States, should develop as a matter of first priority, the Rules of Engagement for any player wanting to provide a component of the EOSC. RoE should be based on the assumption that all Data in the EOSC (in fact all Research Objects) are FAIR (Please note: the FAIR principles do not enforce specific implementation choices beyond checking them on rendering the research objects Findable, Accessible, Interoperable and Reusable). We recommend that the HLEG-EOSC should urgently establish and guide a dedicated group to draft a proposal for the Rules of Engagement in 2016. As the EOSC develops, compliance with the RoE could be implemented as a seal of approval for web resources and services shared and provided on the EOSC.

In call cases, we envisage that implementation will be phased and that there might be immediate, simple RoE for the short-term and then medium-term work towards sustainable long-term take up.

I2.1: Set initial guiding principles to kick-start the initiative as quickly as possible

We offer the following principles to guide the initial, preparatory phase:

1. The EOSC will build on a community-based, lightweight sustainable governance.
2. The EOSC will support existing excellence wherever possible.
3. The EOSC will support both scientists and innovators and will be user driven.
4. While collaboration is needed, this does not mean that scientists and innovators dictate to developers what to build, but that data experts and engineers contribute their respective knowledge and expertise about what is possible for agile developments.

a) scientists and innovators need to coordinate to speak with one voice to provide the developers with a clear and consistent message about their needs.

b) scientists and innovators need to commit to agile development with regular feedback and user testing to avoid expensive solutions that are not fit for purpose.

5. A first cohort of core data experts should be trained immediately to translate the needs for data driven science into technical specifications to be discussed with hard-core data scientists and engineers. This new class of core data experts will also help translate back to the hard-core scientists the technical opportunities and limitations.
6. Based on an internet-type hour-glass model, the EOSC will need community-endorsed, internationally governed and enforceable set of protocols.
7. These protocols should be absolutely minimal, open and transparent so that all scientists, innovators, engineers and service providers understand them, see their value and can adhere to them, even if technology and data formats rapidly develop (as will be the case).
8. These protocols should be tuned down to the very basics of what data and related services need, what they support at the most basic level and only where strictly necessary to make the EOSC work (comparable to TCP/IP, HTTP and HTML for the Internet).
9. These protocols should count for all Research Objects[22] and they should enable the minimal requirements for Research Objects to be widely and effectively (re)-used.
10. The FAIR principles[23] will guide implementations to make research objects Findable, Accessible, Interoperable, ultimately to make them Re-usable and citable.[24]

[22] See glossary
[23] see footnote 11

11. Within the scope of FAIR principles, the standards and protocols should again be restricted to the absolute minimum, to mitigate the risk that future developments will require adaptations of protocols.

12. The complexity of the current data-sharing practices and mechanisms requires gentle rather than restrictive regulation of existing ontologies, especially across domains, with identifier mappings as practiced already in various communities.

13. The EOSC should distinguish domain specific standards and protocols (e.g. Preferred Persistent Identifiers in a discipline) and protocols for concepts and data formats that are of general utility.

a) Domain specific protocols should emerge from community best practices and we endorse a lead role for the existing and future ESFRIs, ERICs and other disciplinary research community federations.

b) Generic protocols should be the responsibility of international scientific organisations both formal and informal (e.g. ORCID for researcher PIDs), ICT standards of e-infrastructure communities, legislators and industrial producers of hardware, software and governments (e.g. visualisation and analytics procedures, generic standards pertaining to software and hardware, single sign-on, authentication, authorisation and protection).

14. The RoE should guide the participation of compliant developers and service providers and of other stakeholders. The authorisation mechanism for providers may be based on self-reporting practices already used in the European Union and on some relevant international schemes.

15. Like protocols, RoE will have to be open, transparent, co-developed with and acceptable by user and provider communities and strict enough to prevent undesired and unacceptable use such as abuse of data, scientific malpractice, unfair pricing, vendor lock-in, monopolisation and unjust exclusion of users.

16. While RoE should be generic, they should focus on specific categories of stakeholders in the EOSC-ecosystem as necessary.

I3: Fund a concerted effort to develop core data expertise in Europe

We recommend a very substantial training initiative in Europe so as to locate, create, maintain and sustain the required core data expertise. This should be a community-based effort led by the major training stakeholders and consortia in the ESFRIs, e-Infrastructures and beyond, such as in national and international training consortia. The aim of this training and education effort should be ambitious:

1. By 2022, to train hundreds of thousands of certified core data experts with a demonstrable effect on ESFRI/e-INFRA activities and collaboration and prospects for long-term sustainability of this critical human resource.

2. Consolidate and further develop assisting material and tools for the construction and review of Data Management Plans (including budgeting for re-use of data) and Data Stewardship plans (including budgeting for data publication and long-term preservation in FAIR status).

3. By 2020, to have in each Member State and for each discipline at least one certified institute to support implementation of Data Stewardship per discipline.

I4: Develop a concrete plan for the architecture of data interoperability of the EOSC

The EC and the relevant constituencies in Member States should have a guiding and governance role in the appointment of implementation bodies and in the oversight and monitoring of RoE compliance. We

[24] https://www.force11.org/group/joint-declaration-data-citation-principles-final

recommend the following concrete actions for the EC in concert with the relevant constituencies in the Member States.

1. Delegate the setting of standards and protocols for domain specific issues to existing national and international constituencies, including ESFRIs, scientific associations/academies; in their absence, physical or functional, stimulate their rapid development via a Roadmap or other suitable guiding instrument.

2. Actively stimulate and support multiple ESFRI-type communities in the same broad domain to collaborate on these issues and collectively set a minimal set of norms for a Preferred Persistent Identifier (PPID) scheme in their domain as well as mappings to other PID

3. Stimulate cross domain collaboration at ESFRI level for more generic semantic types such as people, organisations and geographical locations. In modern science and innovation, research by other researchers are key; we also recommend that the EOSC governance take a stand on the federation of social networking scientific applications and try to engage dedicated person-oriented applications such as Google Scholar, ResearchGate and Academia.edu via the RoE.

4. Specify the role of the Research Data Alliance (RDA)[25]. While many RDA working groups address protocols and other instruments for scientific collaboration based on data, stronger coordination and collaboration with ESFRIs, e-Infrastructures and national infrastructures is needed.

5. Alongside minimal and rigorous protocols, safeguard maximum freedom in the design of protocol-compliant templates, tools, datasets and services.

I5: Install an innovative guided funding scheme for the preparatory phase

To stimulate the required change and innovation and deliver an EOSC by 2020, all measures proposed should not follow traditional and rigid funding schemes of the past for scientific data management - e.g. as a small, unaccounted part of a time-limited and space-bound grant in a discrete national / EC funding schemes - but have the character of a game-changer scheme, modelled on the highly effective DARPA challenges in the US[26]. In addition, well budgeted data stewardship plans should be made mandatory and we expect that on average about 5% of research expenditure should be spent on properly managing and stewarding data (see I5).

The game-changer scheme should have precise aims that are linked to core EOSC areas, rather than being broad and bottom-up topical calls, in the scope of accelerating the development of a fully functional EOSC. Scoping, evaluation, selection and awards of the scheme should be done through mechanisms specifically designed to rapidly prototype, test and reach the goal. The first result will be neither a regular research project nor a long-term infrastructure, rather a proof of concept and implementation vehicle to kick-start the initiative, to provide a point of reference with a clear offer and sustainability model once the scheme is in place. Within the single game-changer scheme, multiple work packages can be funded to test different approaches. The European Structural and Cohesion Fund might be used for this scheme alongside Horizon 2020 funding, thereby stimulating regions to develop broadly applicable components of the EOSC.

We strongly recommend that the Commission set up structured discussions on the game-changer scheme with Member states to start, extend and sustain rapid prototyping make data and tools compatible with FAIR principles.

[25] Obviously, each organisation before getting a formal or informal mandate will need to sign off on the RoE meaning that we need these as a very first deliverable.
[26] http://www.darpa.mil/work-with-us/public/prizes

We suggest to make ***rapid, agile prototyping and reference implementations*** a critical element of the preparatory phase of the EOSC so that already in 2017 exemplar working environments can be implemented in key disciplines in guiding Member States, which can be then be extended in other settings, communities and countries.

Priority actions of the EOSC game-changer scheme

1. Solve socio-technical bottlenecks to reach full operational status of the EOSC. These could include for instance: connectivity issues, security and trust issues, performance issues, standards or format issues and socio-technical hurdles, such as lack of incentives and reward for data publication and sharing. Proposed solutions may be technical, cultural or a combination of both.
2. Develop and sustain core data assets for the EOSC and make them available to the community under well-defined conditions. These may include workflows, analytics programmes and notably existing datasets with FAIR status (including metadata creation).
3. Create and implement a plan for the sustainable provision and funding of core resources of the EOSC in terms of connectivity, scientific data storage and computing.
4. Support the development of one or more publicly available data search engine(s) that find FAIR metadata across trusted EOSC repositories.
5. Develop technologies and approaches to meaningfully measure re-use and scientific impact of Research Objects after their initial publication (e.g. metrics that matter and get recognised).
6. Develop schemes to improve funding and rewards for open data sharing at research performing organisations and funders.
7. Start dedicated efforts to prepare data and research objects for inclusion in the EOSC.
8. Combine single sign-on issues with the connection of social and professional people oriented web applications resulting in a federated identity and credentials for all people in the EOSC.
9. A repository of research vocabularies and a software application to support wider access, re-use and development of vocabularies thereby enhancing interoperability.
10. Urgently develop adequate data stewardship capacity in European Member States.
11. Engage stakeholder communities in a guided and dedicated effort to develop and offer on-line, scalable and re-usable training modules. Not only to train experts but also to drive convergent evolution of standards and practices used.

I6: Make adequate data stewardship mandatory for all research proposals

We recommend that use of present and future instruments in research programming, including Horizon 2020, should only support projects that properly address Data Stewardship issues for open data. Projects that develop isolated and solipsistic data infrastructures, that do not specify FAIR conditions for data, without a sustainability plan and without a clear plan on how the action and its supporting data and human infrastructure, will not contribute to the development of the EOSC vision of a commons as laid out in this plan, and consequently should not be eligible for funding. On the contrary, projects with achievable requirements, that are data-multidisciplinary, that properly address post-project sustainability or otherwise advance the common aims of the EOSC should be streamlined for funding in the EOSC game-changer scheme.

I7: Install an executive team to deal with international coherence of the EOSC

The EOSC should not develop in splendid European isolation. Sister initiatives in other major scientifically leading regions such as the USA, Australia, and in emerging scientific regions should be actively engaged. We recommend installing a specific EC-endorsed team or organisation to deal with these global issues.

I8: Establish an executive team to deal with the early preparatory phase of the EOSC

In order to kick-start the EOSC, a number of focused task forces should be set up to address the following issues:

Define the RoE of the EOSC and practically test them.
Create and foster cross-domain working collaboration including ESFRIs and e-Infrastructures and address cross disciplinary join-up.
Define training needs for the necessary data expertise and draw models for the necessary training infrastructure.
Develop a governance plan for the EOSC, as lightweight and inclusive as possible.
Establish minimal technical standards for the EOSC and plan for their long-term maintenance and compliance.
Establish guidance and oversight mechanisms for the EOSC fame changer scheme as a development outside of regular programming calls.
Liaise with Member States to establish plans for certified institutes for data expertise and stewardship and to help building capacity for certification.

Glossary of terms

1. **Data:** When the term data is used in this report we refer to digital research objects in a broad sense, including regular research data and also meta-data, the associated services and workflows, analytics algorithms and all other data-related instruments that modern scientific research uses.

2. **Data Stewardship:** The entire process that deals responsibly with one's own and other peoples data throughout and after the scientific discovery process.

3. **Commons**: we consider the EOSC as the European contribution to a global scientific commons. When we use this term we refer to the concept of a common public good that allows data publication, stewardship and re-use. We emphasise that public as well as private service providers can participate in this commons.

4. **Data Experts**: this term refers to a distinct and largely novel class of research professional. They are not traditional core computer or data scientist, but embedded data specialists that are able to support domain specific researcher throughout the entire knowledge discovery cycle. They typically do not end up with high impact factors in traditional systems but should become indispensable core partners in any modern data driven research team with a solid perspective.

5. **Innovation**: we explicitly refer to the transition from scientific insights to novel societal assets and not to the creative knowledge generation process of scientific activity.

6. **Machine Actionable**: when this term is used, we explicitly refer to machine readable and executable data or meta-data. The FAIR principles emphasise that need as well. It is pertinent in data driven science that machines can operate on data as relatively independent agents.

7. **ESFRI**: this term formally refers to the European Strategy Forum on Research Infrastructures. In this report the term is loosely used to refer to large, domain oriented international research infrastructures, regardless of whether they are funded under the European or National ESFRI Roadmaps.

8. **e-Infrastructures**: this term is used to refer in a broader sense to all ICT-related infrastructures supporting ESFRIS or research consortia or individual research groups, regardless of whether they are funded under the CONNECT scheme, nationally or locally.

9. **Rules of Engagement (RoE)**: this term is used to refer to the rules of engagement and conduct that need to be developed in the preparatory phase of the EOSC.