High Level Expert Group for the European Open Science Cloud – INITIAL DRAFT REPORT and Recommendations

Version: 09/01/15 Semi-Final for discussion at 13 January 2016

Vision

The year is 2030. Open Science[1] has become a reality and is offering a whole range of new, unlimited opportunities for knowledge sharing and discovery worldwide. Scientists, research institutions, publishers, public and private research funders, students and education professionals as well as companies and citizens from around the globe are sharing a largely open, virtual research environment.

Open source and data communities and scientists, publishing companies and the high-tech industry have pushed the EU and UNESCO to develop a 'Commons' for Research and Innovation with community endorsed and open research standards, establishing a virtual learning gateway, and offering effective access to all scientific data as well as to all publicly funded research tools and methods.

The OECD (which now includes Brazil, India, China and Russia), as well as many countries from Africa, Asia and Latin America have adopted these new standards, allowing users to share a common platform to exchange knowledge at a global scale and across traditional disciplinary, geographical and social barriers. Social Machines are an integral part of this virtual research environment.

High-tech startups and small public-private partnerships have spread across the globe to become the service providers of this new digital science, innovation and learning network, empowering researchers, educators and students worldwide to share knowledge by using the best available technology.

Most of our 8.5 billion citizens contribute constantly to the data stream from their blood-circulating Nanobots and their wearables as well as by active annotation and sharing efforts. Social Machines process these data in near-real time and self-learning, fluid graphs will make consistent patterns emerge and people will derive 'actionable knowledge' from these Social Machine environments with a speed that is inconceivable today. Broadly accessible and open, high quality and crowd-sourced science, focusing on the grand societal challenges of our time, shapes the daily life of a new generation of citizens.

(adapted from 'Open Science in 2030): http://ec.europa.eu/research/openscience/index.cfm

This report and the associated recommendations address the infrastructural consequences of the transition to Open Science and Open Innovation, from data-sparse to data-rich and data driven enterprises. A more open and participatory way of global knowledge sharing and application is enabled by the wealth of data we create in- and outside the scientific realm. How does Europe, with its long scientific tradition, optimally support this transition, avoid its potential downsides and connect to the rest of the globe to serve research?

[1] **See for background and policy document: http://ec.europa.eu/research/openscience/index.cfm**

Executive Summary

The perceived **European Open Science Cloud (EOSC)** aims to accelerate and support the current transition to more effective Open Science[2] and Open Innovation[3] in a Digital Single Market[4]. It should enable trusted access to services, systems and the **re-use of shared data** across disciplinary, social and geographical borders. The term 'Cloud' is a **metaphor** to help convey the idea of seamlessness and a '**Commons**'. The 'EOSC' is approached in this report as a **federated environment**, composed of elements in the Member States, with **minimal international guidance and governance** and **maximum freedom to implement.** The EOSC is indeed European, but it should also be a **globally interoperable** and **accessible** infrastructure. It includes the required **human expertise**, **resources, standards, best practices** and underpinning **infrastructures**. An important aspect of the EOSC is therefore **systematic and professional data management** and long term **data stewardship**. However, data stewardship is not a goal in itself and therefore the final realm of the EOSC is the **frontier-science and innovation process** in Europe.

Challenges and observations:

- The majority of the identified challenges to reach a functional EOSC is **social** rather than **technical**
- The major challenge is **not the size of Data** *per se*, but in particular **complex data** and analytics across domains.
- There is an alarming shortage of **data experts** globally and also in the European Union
- This is partly based on an **archaic system of rewards** and **funding** of science and innovation
- The lack of core 'intermediary expertise' has also created a '**valley of death**' between **(e-)infrastructure providers** on the one hand and d**omain specialists** on the other.
- The short funding cycles of **core research infrastructures** are **not fit for purpose**
- The **fragmentation** (even now that the ESFRI scheme is highly successful) between domains causes **repetitive** and **isolated** solutions
- The ever larger distributed data sets increasingly **do not move** (for sheer **size** or for **privacy** reasons) and centralised HPC is therefore **insufficient** to support the critically **federated and distributed meta-analysis an learning**.
- Notwithstanding, all these current hurdles and challenges, the **major components** needed to create a **first generation EOSC** are largely 'there' but '**lost in fragmentation**' and spread over 28 Member States.

Key factors for the effective support of data driven Open Science and Innovation

- 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
- Innovative, fit for purpose funding schemes are needed to support sustainable underpinning infrastructures
- Core data experts need to be trained and their career perspective significantly improved
- Cross-disciplinary collaboration requires specific measures in terms of review, funding and infrastructure
- The transition from scientific insights towards societal innovation needs a dedicated support policy
- The EOSC needs to be developed as the data infrastructure Commons as an eco-system of infrastructures
- Key Performance Indicators should be developed for the EOSC
- The EOSC should where possible enable automation of data processing and thus machine actionability is key.

Specific recommendations to the Commission for a Preparatory Phase.

P1: Take immediate, affirmative action 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 supporting Internet based protocols and applications
G1: Aim at the lightest possible, internationally effective governance
G2: Guidance only where guidance is due
G3: Define Rules of Engagement for formal participation in the EOSC
G4: Federate the Gems across Member States
I1: Turn this report into an EC approved White Paper to guide EOSC initiative
I2: Develop, Endorse and implement a Rules of Engagement scheme
I3: Fund a concentrated effort to locate and develop Data Expertise in Europe
I4: Install a highly innovative guided funding scheme for the preparatory phase
I5: Make adequate data stewardship mandatory for all research proposals
I6: Install an executive team to deal with international coherence of the EOSC
I7: Install an executive team to deal with the [1st Q] preparatory phase of the EOSC

[2] See for background and policy document: http://ec.europa.eu/research/openscience/index.cfm

[3] See speech of Carlos Moedas: http://europa.eu/rapid/press-release_SPEECH-15-5243_en.htm

[4] http://ec.europa.eu/priorities/digital-single-market/

The European Open Science Cloud?… some nuances and definitions

As a first step towards the '2030 scenario'; 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 is able to fully contribute 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 the Global Research Data Commons.

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

It specifically aims to accelerate the transition to more effective Open Science[5] and Open Innovation[6] in a Digital Single Market[7] by removing the technical, legislative and human barriers to research data and tools re-use, and by supporting access to services, systems and the flow of data across disciplinary, social and geographical borders. The term, European-Open-Science-Cloud therefore first requires some reflection as the term may infer some incorrect associations and boundaries that need to be clarified, and in fact the term 'Cloud' is a metaphor to help convey the idea of seamlessness and a 'Commons'.

- '**European'**: we recognise that research and innovation are global, and as such science is a global issue. The (E)OSC can therefore not be built exclusively in and for Europe. Serious efforts are needed to ensure coordinated action with other geographical regions. However, 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[8]. 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'[9]
- **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 will support not only data driven scientific research but should also facilitate 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 also can be mis-interpreted and indicate that the EOSC is mostly about 'hard ICT infrastructure' and much less about a Commons of software, standards and expertise related to data-driven science and innovation.

---

[5] See for background and policy document: http://ec.europa.eu/research/openscience/index.cfm

[6] See speech of Carlos Moedas: http://europa.eu/rapid/press-release_SPEECH-15-5243_en.htm

[7] http://ec.europa.eu/priorities/digital-single-market/

[8] although scientists may perceive things that 'other people paid for' to be 'free' for them and ready to turn against any commercial approaches even if they are demonstrably better then 'free' alternatives.

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

Why do we need a European Open Science Cloud?

The EOSC is a need emerging from science in transition[10]. The desired 'EOSC' is indeed European, but it should also be a interoperable with the Global Research Data Commons and 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, 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 modern data management and long term data stewardship.

Europe currently enjoys a long tradition and a relatively healthy research infrastructure, served via domain specific research infrastructures and cross-domain ICT infrastructures, as well as disciplinary and cross disciplinary collaborations and services. Alongside this, 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 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 the necessary expertise and strong collaboration.

Science itself is in an unprecedented phase of transition driven by the power of 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 significant transition as well. Many of these are rooted deeply in the scientific community and in the support structures of research and they appear to be quite resilient to this quantum leap.This fundamental shift nevertheless is required to match the potential to generate ever increasing amounts of data and to turn them into knowledge as the fuel for innovation and also to meet global challenges. This 'phase transition', after considerable consultation and debate was coined by the EC as a transition to 'Open Science'[11]. The EOSC is a fundamental environment that needs to be realised to underpin and enable this transition.

[10] the following points are all supported by a long range of recent policy and position papers. These will be listed in Annex 1 of the report, but for readability purposes we will keep the number of footnotes and in-text citations as minimal as possible. A summary of the most 'controversial statements' in this advice will be given in Annex1, with the major supporting earlier policy and position papers as references

[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.

The key aspects of modern ‘Open Science’ are:

- **New modes of scholarly communication;** Scholarly communication, which has been dominated by narrative and verbal means of delivery for centuries, needs to move much more 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 close to zero support for publishing of new data and for tool sharing; the development and reward of data related expertise; data stewardship and (re-) analysis to support the final aim of science: knowledge discovery.

- **Train and sustain core data experts,** especially in academia where they are most severely undervalued. A lack of data related core expertise may well be among the risks leading to Europe loosing its leading position in science. As argued before, there is a market failure as the formal reward mechanisms within the research system are biased towards the traditional research paper and journal publication; new forms of output and data and software 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:** Not only is cross disciplinary collaboration critically needed (yet hampered by current policies and practices) but also, scientists will, in open science, increasingly use valuable raw and curated data resources, as well as analytics tools from disciplines other than their own. However, currently, ’other peoples' data' is already notoriously difficult to discover even within one's own discipline. Discovering relevant (other peoples) 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'. But, how would a health researcher know about a valuable datasets 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 related lack of 'data search engines' researchers cannot be blamed for re-inventing a new wheel. We can assume that this is amplified further across disciplines.

- **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: SME's is relatively weak. Multidisciplinary research and innovation projects and 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.

- **An eco-system of infrastructures:** it may seem counter-intuitive at first glance but the challenges of ever bigger data can no longer be simply solved by ever bigger infrastructure. Next to advanced computer science, that hopefully will bring us innovative forms of computing and storage, 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 as well as the 'Moores law' increase in storage and computing power, many comprehensive datasets are simply too big to move. Increasingly, data are so privacy sensitive that legislation effectively precludes their physical move outside the safe environment in which they reside.Therefore, relatively 'featherlight' workflows (e.g. process virtual machines) containing parallel and distributed analytics algorithms will increasingly 'visit data where they reside’, with supporting reference data and transporting only ‘conclusions’ outside the safe data vault. 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 'integrated'.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. Therefore in some senses, machines have become essential research assistants, both in data generation and in data processing and analytics. The *'excel age'* is definitively over. Data formatting, terminology/identifier mappings and provenance must therefore be optimally organised in order to support the machine processing as well as the human knowledge extraction from data. 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[12] is therefore a crucial consideration in modern data publishing.

A Key challenge: Modern science drives two very different communities and cultures together

In an earlier transition of scientific research from a largely individual, elite and intellectual activity to a mainstream, largely institute based activity, we introduced a new profession: the research analyst, and in the laboratories of the leading academics these people soon became indispensable and highly recognised research professionals, co-publishing and being involved in all aspects of the research and experimental cycle. When data generating machines became mainstream and consequently high-throughput data generation boomed, the experts who knew how to operate these data generating machines only gained in importance.
But what about the *data analytics* machines? When computers became natural key 'partners in research', a peculiar trend developed. Over the last decade key actors in modern research (computer and data specialists) are not given the same personal credit in the scientific research process as wet lab analytical people got in the past. Why?
The scientific cultures from which these experts come have different reward systems and incentives[13], different jargon and very different skill sets. These cultural differences are resulting in unnecessary mis-matches and in the alarming loss of crucial data related skills in research; we have what perhaps can be characterised as an unhelpful divide between researchers and those that support research with data processing and software. As a consequence, these two communities that are essential to Open Science have not closely co-evolved. Often the front-line ICT developments take place rather independently from 'day to day experimental or social science'. In contrast to other lab-equipment, experimental scientists frequently misjudge ICT infrastructures as supporting infrastructure that can be relatively easily purpose built and they underestimate the complexity and the need for professionalism. (the added value of working together is not obvious and visible, recognition)
In addition, support for data generating scientific activity and the support for the underpinning research infrastructure has traditionally also been separated, both in many Member States and at the EC level; this may have further aggravated this separation of worlds. To a large extent this is an understandable divide but it has also contributed to the sub-optimal communication and collaboration between the top-tier ICT experts and the top-tier experimental and social scientists. Where professionals have been able to bridge the divide and have effectively collaborated major advances have been made. Most researchers are still frequently struggling forward with severely suboptimal solutions, sometimes out of sheer ignorance of what is available, but often because actual collaboration with the 'computer scientists and engineers' takes time and is not easy.

[12] SDATA-15-00190: Wilkinson et al 2016: Wilkinson et al: FAIR Data: Guiding Principles for Scientific Data Management and Stewardship, in press; Nature Scientific Data

[13] for example: 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)

This all results in the unfortunate situation that the Dunning-Kruger effect[14] is very apparent in the clash of cultures between the current main stakeholders in ICT/e-INFRA and experimental Science. The complexity, the cost and the intellectual challenges in the other domain or discipline are systematically underestimated and undervalued (both ways). This hampers the same 'user driven' and 'expert enabled' co-development of core ICT infrastructure that is required so that research can meet its full potential.

Moreover, very little incentive exists in the reward system for 'customer support', and support for 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 and consuming new data types. Once e-infrastructure and tools become 'commodities' many scientists do not see the logic of co-authorship on scientific publications but rather they 'acknowledge' data experts for 'analysing the data'. Agile co-development with continuous power-user feed back and rigorous testing of prototypes is not only precluded by the cultural differences, but also because users do not always conceive of the possibilities the latest developments in data science enable. This is certainly contributing to the lack of data scientists that venture out from classical computer or data science departments into other scientific fields.

Data Expertise is lacking particularly in the EU

There is a pressing requirement with regards to the necessary data expertise that will support the aims of the EOSC. It became clear and has been reflected in nearly all contributions so far that there is a major hole in the EOSC planning if we do not repair the significant lack of 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 breed of colleagues* with core scientific professional competencies. Core Data Experts are neither 'computer savvy experimentalists' (although the latter also need to be educated to the point where they hire, support and respect Data Experts) nor are they hard core data or computer scientists or software engineers. They should though be proficient enough in the domain where they work to be routinely consulted in the group 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 part of the picture (including data re-use and stewardship planning and proper budgeting) and the proper capturing of new data (formats, metadata richness, standards, provenance, publishing, linking and analysis). This expertise is rare and the people with these skills are often attracted to industry or outside Europe where they are more respected and valued.

The alarming lack of reproducibility of current published research as widely publicised, which together with scientific fraud does enormous damage to the reputation of science. This problem is partly due to the lack of deep and rigorous knowledge on how to render data and the associated tools in the format that allows others to reproduce the results. The good news is that not necessarily all conclusions in the literature for which the results can not be easily reproduced elsewhere are wrong, but reproducibility and early detection of fraud-signals and re-use will increase as a result of core data processing and analysis expertise. The number of people with these skills needed to effectively operate the EOSC is likely exceed 500,000 within a decade[15] (references needed HLEGx). We believe that implementation of the EOSC needs to include steps to help train, retain and recognise this expertise, in order to support 1.7 million scientists[16] and over 70 million people working in innovation. The success of the EOSC depends upon it.

[14] Unskilled and unaware of it: People tend to hold overly favourable views of their abilities in many social and intellectual domains. The authors suggest that this overestimation occurs, in part, because people who are unskilled in these domains suffer a dual burden: Not only do these people reach erroneous conclusions and make unfortunate choices, but their incompetence robs them of the metacognitive ability to realise it. Original article in: Journal of Personality and Social Psychology, Vol 77(6), Dec 1999, 1121-1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121

[15] (references needed HLEGx)

[16] define scientist

How will the European Open Science Cloud be realised?

Policy, Governance, First stage implementation and guiding principles

If one consensus arose from the consultations, the policy papers and the debates with the stakeholders, it is that the EOSC should *not* be a new major, localised and centrally governed initiative. We believe that the 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 a Global Research Data Commons.

There is a clear analogy with the early days of the current 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, URI's and HTML drove its major application domain, the largely textual WWW. This combination of 'authorisation', key support by a major leading agency (in this case NSF) and a dedicated community (W3C) setting absolutely minimal standards allowed virtually everyone to start building standard-compliant tools and services in the ecosystem. The Internet still has no 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[17].
In practice, the standards for the Internet were so minimal and rigorous that even after 25 years of overwhelming growth and development up to the now ubiquitous smartphone, the basic standards are still essentially the same. Only recently was there a move to IPV6, which is predominantly to allow a much broader range of IP addresses and it does not fundamentally change any of the other features. When XML and RDF developed in a first attempt to develop schema free and self-defining components in the internet, nothing fundamentally changed.

That is what we need to achieve again and we propose to stay as closely to the lessons learned and the choices made for the narrative applications on the internet and the early stage semantic web.

At the European policy- and organisational level the EOSC should take a similar approach 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 to the ESFRI approach. For example we cannot afford a preparatory phase of many years, as the need of many disciplines for an early functional EOSC is very clear from all position papers most elements have been judged as 'being there but hidden in fragmentation. We therefore need to commence defragmentation actions immediately, including the setting up of light and appropriate guidance and governance structures and prototyping for new solutions that are needed during the preparatory phase. The recommendations that follow are mainly for this proposed preparatory phase. **Recommendations** will be 'Policy' (P), Governance (G), and Implementation (I) related.

[17] Hart, Strawn, A Brief History of NSF and the Internet, August 2003, https://www.nsf.gov/od/lpa/news/03/fsnsf_internet.htm

Recommendations of the High Level Expert Group

P1: Take immediate, affirmative action 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 others data and tools for research, innovation and educational purposes under well defined, secure and trusted conditions, supported by a sustainable and just value for money model.

P2: Close discussions about the 'perceived need'

Although the EOSC ultimately will need to involve all ESFRIs, all e-INFRAS, private sector and all 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 some 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 annex 1), 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 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 are mainly cultural and reasonably well defined. We therefore believe that many of the actions we propose for the preparatory phase of the EOSC can be significantly progressed or completed by the end of 2017.

P4: Frame the EOSC as the EU contribution to a Global Research Data Commons, with open protocols.

The current internet application domain seems to have avoided the complete 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. Following an approach to the EOSC which is based on minimal rigorous standards will prevent costly and time consuming exercises to decide who has the best solutions. Instead this approach will allow participation from all stakeholders, including e-infrastructure providers, Member States, research institutes and businesses. All providers, public and private, can start implementing prototype applications for the Global Research Data Commons on the day they have the minimal standards and the minimal rules of engagement.

G1: Aim at the lightest possible, internationally effective governance

Given the urgency and the number of stakeholders and participants required to realise the EOSC we believe a strictly governed, new infrastructure built 'somewhere' or even 'everywhere' is not the right model for the EOSC to be a success. Instead a more inclusive, flexible, transparent and less centralised approach is required, that also enables effective global collaboration.

G2: Guidance only where guidance is due

Of course while we advocate lightweight governance it does not mean we need absolutely no regulation. The current 'standards jungle' needs to be actively regulated and for example some major players (both public and private) may claim an unjust and counterproductive share in the EOSC. The EOSC will obviously have a myriad of small and very large players, as in the current narrative Web, but it is perceived (even more so than the textual Internet applications) as a 'public good' where citizens, researchers and innovators need to use each others data in a trusted affordable and sustainable environment.

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

All players, public and private, European and non-European should be able to provide data and/or services in the EOSC and the associated expert infrastructure.
The EOSC will be guided and governed by a minimal set of rigorously applied and enforced standards, so called 'Rules of Engagement' (RoE). These RoE will be used to 'brand' (a subset of) providers in the EOSC as 'trustworthy and compliant with the RoE'. It should be clear that non-EOSC approved players are free to explore any role in the EOSC they wish, even if they do not adhere to the RoE. They will just not be able to brand their services as **['EOSC-approved']**

G4: Federate the Gems

Based on the consensus that most foundational building blocks of the Global Research Data Commons are operational somewhere, but that they operate in domain, geographical and funding scheme silos, we recommend that early actions concentrate on what might be termed the '**federation of the gems**'. The optimal engagement of the e-INFRA communities, the ESFRI communities, the building blocks of those in individual Member States, but also the wealth of small and large industrial players in Europe should be stimulated. All partners and stakeholders that adhere to standards and sign off on the RoE should be eligible.

I1: Turn this report into an EC approved White Paper to guide EOSC initiative

This report can be the basis for a formal white paper to be endorsed by Member States and the EC and which can serve as a guiding document for actual developments and implementations in the Member States and H2020, as well as a discussion basis for further international consensus building and collaboration.

I2: Develop, Endorse and implement a Rules of Engagement scheme

The Commission, in close collaboration with appropriate stakeholders in the Member States should develop as a matter of first priority, the Rules of Engagement for any player that wants to provide a component of the EOSC. Obviously, these should include that all Data (Research Objects) in the EOSC are considered FAIR (again, this principle *does not enforce implementation choices* beyond checking them on rendering the Objects Findable, Accessible, Interoperable and Reusable) We propose that the HLEG-EOSC should install and guide a dedicated group to draft a proposal for the Rules of Engagement in the first quarter of 2016.

As the EOSC develops the compliance to the 'RoE could be implemented in the form of a 'EOSC' seal of approval, which providers can put on their web resources, services and communication materials.

We recommend to further develop the following rough guiding principles' for the preparatory (and the early parallel exemplar implementation phase):

1. The EOSC builds on community emergent, guided but lightweight governance
2. The EOSC engagement scheme will support existing excellence wherever possible
3. The EOSC supports Scientists and Innovators, and will be user driven.
4. This means collaboration is needed, and importantly it does not mean that 'scientists just tell engineers what to build', but that data experts and engineers contribute their knowledge and expertise about what is possible to the agile developments.
5. Researchers and Innovators need to coordinate to speak with one voice to provide the data experts and engineers with a clear and consistent message about the needs
6. Researchers and Innovators need to commit to agile development with regular feed back and user testing to avoid solutions that are not fit for purpose
7. We need to train intermediary experts who translate the needs for data driven science into technical specifications to be discussed with the hard core data scientists and engineers.
8. This 'new breed' of core data experts will also be able to translate (back) technical opportunities and limitations to the hard core domain scientists.
9. Based on the highly successful and sustainable principles of the internet and the so called 'hour glass model', the EOSC will need a community endorsed, but internationally governed and enforced set of standards

10. These standards should be absolutely minimal, completely 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)
11. Therefore these standards should be entirely tuned down to the very basics of what data and related services are, what they support at the most basic level and what 'can absolutely not be avoided' to make the EOSC work (comparable to TCP/IP, URI, HTTP and HTML for the Internet of Hypertext.
12. These standards should count for all Research Objects[18] and they should enable the minimal requirements for Research Objects to be widely and effectively (re)-used.
13. The FAIR guiding principles (ref: in themselves not a 'standard') will be leading as implementations following these guiding principles will render research objects Findable, Accessible, Interoperable in order to reach the final aim and make them 'Re-usable'
14. Within the scope of FAIR guiding principles, the actual standards should again be restricted to the absolute conceivable minimum, to mitigate the risk that future developments will require adaptations of standards.
15. The complexity of the current standards reality requires to avoid inhibitory regulations on conceptual ontological references, especially across domains. The Open PHACTS (IMI) and ELIXIR proven approach of a Preferred Persistent Identifier, accompanied with Identity Mapping and Resolution Services, effectively work and preclude endless discussions on 'pet' identifier schemes. The solution is to (a) support all appropriate systems and (b) put the responsibility of PID>PPID mapping to the user of the alternative PID.
16. We need to distinguish **domain specific standards** such as **Preferred** Persistent Identifiers for all concepts referred to in a discipline and highly specific data formats. The domain specific standards should emerge from domain community best practices and we propose a strong role for the existing and future ESFRIS, ERICs and other topical Research Community Federations.
17. The **generic standards** are mainly the realm of international organisations (for instance ORCID for researcher PID's) and ICT standards of the e-INFRA communities, legislators and industrial producers of hardware, software and governments and include for instance visualisation and analytics procedures, and generic standards pertaining to software and hardware standards, single sign-on, authentication, authorisation and protection.
18. The Rules of Engagement themselves should serve a guided and controlled participation of compliant providers. The authorization scheme could be based on the self-reporting scheme already practiced in the CE implementation and in the USA [reference]
19. The EOSC is open to everyone, but formal participation (other than as an occasional consumer) will work on the basis of Rules of Engagement (RoE)
20. Like the standards, the RoE will have to be open, transparent, co-developed with and acceptable by all targeted user and provider communities and just strict enough to prevent undesired and unacceptable use such as abuse of data, exorbitant pricing, vendor lock in, monopolisation and unjust exclusion of less privileged users. (See HLEG011 for exemplar rules engagement)
21. The RoE can be developed generically and in some cases for specific categories of stakeholders in the EOCS-ecosystem.

I3: Develop a concrete plan for the light weight governance of the EOSC

The EC and the relevant bodies in the Member States will need to have a guiding and governance role in the appointment of standards bodies and the oversight and policing of the RoE compliance. Therefore we recommend the following concrete actions for the EC in concert with the relevant constituencies in the Member States

1. Delegate the setting of standards for science-domain specific issues to (preferably existing) national and international 'ESFRI-type' constituencies and in their absence for a given domain stimulate their rapid development via a roadmap.

[18] will be defined earlier in a 'definitions' section

2. These 'standards', include for instance [19]' formats, metadata schemes, controlled vocabularies, ontologies and best practices and they MUST be kept light so they are conducive to the rapid development of the EOSC.
3. 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 something to be called and used as an '**Preferred** Persistent Identifier' as well as mappings to other PIDs.
4. Stimulate with some urgency the cross domain collaboration at 'ESFRI' level for more generic semantic types such as people (e.g. ORCID), organisations (e.g. VIVO, ISNI) and geographical locations.
5. We propose to define a specific global role for the cross-domain Research Data Alliance[20]. In the RDA there are already many working groups that address these kind of issues, but stronger coordination and collaboration with ESFRIs and e-INFRA's is needed.
6. Next to the insurance that standards are minimal and rigorously kept, the safeguarding of 'maximum freedom' in the design of standard compliant templates, tools, datasets and services may need a light 'governance body' as well.

I3: Fund a concentrated effort to locate and develop Data Expertise in Europe

We recommend a very substantial training initiative in Europe so to create, maintain and sustain the required core data experts. Again this should be a community based effort lead by the major training stakeholders and consortia in the ESFRIS and the e-INFRAs and beyond, such as in international training consortia.

The aim of this training and education effort should be ambitious:

1. By 2020, we have trained hundreds of thousands of core data experts at all appropriate levels with a demonstrable effect on ESFRI/e-INFRA effective collaboration and prospects for long-term sustainability of this critical resource.
2. Training and capacity building efforts should include consolidation and further development of assisting material and tools for the construction and review of Data Management plans (including budgeting for re-use of other data) and Data Stewardship plans (including budgeting for data publication and long-term preservation in FAIR status).
3. By 2020 we have in each Member State and for each discipline minimally one 'certified institute that can review and approved DS plans.

I4: Install a highly innovative guided funding scheme for the preparatory phase

To stimulate the required change and innovation, any measures discussed here should NOT follow traditional and rigid funding schemes of the past, but have the character of 'challenges', modelled on the highly effective 'DARPA' challenges in the United States.

The challenges should define precise aims, rather than being broad topical calls, that need to be reached in the scope of accelerating the development of a fully functional EOSC. Selection, award and evaluation of the winner(s) of these challenges should be done by mechanisms specifically designed to reach the goal: These are NOT regular research projects, nor are they long term infrastructures, rather they are 'proof of concept studies and implementation studies' with the requirement to develop a clear offer and sustainability model once the challenge is completed. Some of these could take the form of post hoc 'Prizes' and even involve Crowdfunding. Multiple challengers can be funded to test different approaches.The Regional Infrastructure funds may potentially be used for this innovative scheme, thereby also stimulating developing regions in the Union to develop broadly applicable components of the EOSC. We strongly recommend that the Commission set up structural discussions with the Member states to start, extend and sustain such a 'rapid prototyping' and 'data FAIRification' support scheme. We propose to make ***rapid, agile prototyping and reference implementations*** a critical element of the preparatory phase so that already in late 2016 and ramping up in 2017 exemplar working environments can be implemented in 'guiding disciplines' in 'guiding Member States', which can be replicated in other settings, communities and countries.

---

[19] **Preferred** Persistent Identifiers are PID's that have been designated by the mandated communities to be preferentially used. If other communities or individuals want to use another PID that can be accommodated under the EOSC, but the responsibility to map the alternative PID to the PPID is with the alternative PID user.

[20] 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 (good news is that they are 'almost ready in HLEG011).

4. Potential 'EOSC challenge' categories:

1. One particularly urgent action in this scheme will be the development of adequate data stewardship capacity in European Member States [Annex 1]
2. We need to engage the stakeholders 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.
3. We need to solve perceived or encountered 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, but also socio-technical hurdles, such as lack of incentives and reward for data publication and sharing. Proposed solutions could also be technical or social 'engineering' or a combination of both.
4. We need to develop and sustain core (data) assets for the EOSC and offer them under well defined conditions to the community. These could include workflows, analytics programmes, but notably also valuable existing datasets in FAIR status (including metadata creation).
5. We need to create and implement a plan for the sustainable provision and funding of so called 'core resources' as part and parcel of the EOSC
6. We need to support the development of one or more publicly available data search engine(s) that find FAIR metadata across trusted EOSC repositories
7. Technology and approaches to meaningfully measure re-use and scientific impact of Research Objects after their initial publication (*altmetrics that matter* and get *recognised*)
8. Schemes to improve funding and reward traditions at research performing organisations and funders
9. We recommend to start dedicated efforts to prepare data and other RO's for participation and availability in the EOSC.
10. We need a specific and well thought effort to 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.
11. A research vocabulary repository and portal to support wider access, re-use and development of vocabularies thereby enhancing interoperability

I5: Make adequate data stewardship mandatory for all research proposals

1. We recommend that use of future and present instruments in H2020 should only support proposals that properly address Data Stewardship issues.
2. We recommend that only proposals that develop 'infrastructure' with a sustainability plan and with a clear plan on how the proposed infrastructure will persistent and contribute to the development of the 'EOSC' vision as laid out in this plan will be eligible for funding.
3. We need to specifically support, with achievable requirements, projects that work in multidisciplinary and where appropriate in public-private consortia to address post project sustainability.

I6: 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 developing regions should be taken into account and actively engaged.
We recommend to install a specific, mandated team at the EC level to deal with these global issues.
In modern science and innovation, research by other researchers are key; we recommend that the EOSC governance will also take a stand on the federation of social networking applications, as well as the dedicated people oriented applications, such as Google Scholar, ResearchGate and academia.edu and try to engage them via the RoE.

I7: Install an executive team to deal with the early preparatory phase of the EOSC

TASK FORCES:

- cross-domain/ESFRI (including cross disciplinary and e-INFRA/ESFRI) (clusters)
- Training infrastructure
- A light-weight governance plan for the EOSC
- Technical minimal standards
- Worked out and practice-tested Rules of Engagement (CE-study-connection?)
- A guidance/governance for the 'EOSC challenges programme' (note" this is OUTSIDE the regular planned calls)
- A team to visit member states on capacity building for certified institutes

ANNEX 1 below (please review)

Annex 1 to HLEG-EOSC Report

The EOSC issue of 'reviewing' DM, DS and DR plans (together referred to as DMP's)

There was an intensive discussion on the future 'mandatory' DMP's wherein costs incurred for data management, curation, publication and 're-use of other people's data' would be eligible for funding.
An overall average of 5% of the total project costs was seen as a reasonable estimate for the amount of data quality and re-use supporting funding that this would 'allocate' in future H2020 but also increasingly in MS data intensive research projects.

The proper review (and post award monitoring) of such DMP's in proposals was debated and the options suggested by several funding agencies who are early movers in this area were considered:

1. Review is done at a 'second' stage (maybe even as part of the contract negotiation of already provisionally 'accepted' proposals').
2. Review is done as an integral part of the overall proposal review process.
3. The review of DMP's is done separately from the overall review process and by the time the proposal is submitted the DMP is 'signed off' by some sort of recognised 'authority'.

It became clear that option 3 was the most actionable option.

In option 1, a major issue could become that proponents state that 'their DMP is in order' in the overall proposal and in the 'negotiation phase' some serious issues may surface that either severly delay or complicate the final negotiation process, or even lead to 'rejection' in the second stage. Regardless of outcome of this process, it would place a lot of burden on the staff of the funding agency and the proponents.

In option 2, each panel should in principle contain sufficient DMP expertise and this would draw very heavily on the already scarce expertise about this critical issue in the research community.
It would also bear the risk of lengthy discussions about the validity of the DMP in a panel where the majority of the participants do not have the appropriate expertise to contribute meaningfully to that discussion.

Option 3 brings with it mitigation of the risks expressed for option 1 and 2 as well as a significant effect on 'adequate assistance' for proponents in the development of proper DMP's and of domain specific and generic capacity building in the member states in this crucial field.

(a) Most experimental and social researchers are not (yet) familiar with the many, and complex aspects of good Data Stewardship practices in data-driven science. They may need assistance with the development of adequate DMP's and this assistance should preferably available within the institution and if not available internally, then as close as possible to the investigator's location and with the lowest possible barrier to entry. The latter could also mean that domain expertise should be present in the assisting external facility.
(b) The co-development and 'sign off' of DMP's for research proposals by professional and if at al possible 'certified' institutes (or departments in institutes) would relieve national and international review boards and panels from the burden to 'judge' the DMP in detail. In fact a statement very similar to what we accept for years in 'ethical paragraphs' (a statement that ensures proper review by the institutional ethical board has been conducted and led to approval of the ethical considerations) could be considered. A question of the the nature: 'has your DMP (only briefly described in the proposal) been reviewed and approved by a recognised authority in your institution/country? (yes/no) would have a number of distinct expected effects:
    (a) It would prompt research institutions to safeguard their competitive edge by training and installing (with good HR perspectives) appropriate expertise and support staff in their institution (or rely on trusted third parties) to equip their research staff with the proper consultancy, review and budgeting assistance.
    (b) It would stimulate countries and research consortia to develop, install and sustain professional institutes, groups and/or networks that can offer this service locally and where need close to domain specific as well to e-INFRA specific knowledge and expertise.

(c) It would enable funders (including the EC) to develop a catalogue of 'approved' DS entities in the member states that would be 'allowed to sign off on DMP's' so that this a 'tickbox issue' at scientific merit and feasibility review.
(d) This approach would thus have a major stimulating impact on training and capacity building in the member states on this crucial expertise category in data driven science and therefore boost the competitiveness of EU research in general, especially if dedicated and significant training and capacity building funds would be made available by the EC and the collaborating MS.

(c) We propose to recommend option 3 and make the rapid development of this expertise and the 'certification' actions the topic of one of the 'challenge funds' associated with the rapid development of the EOSC.
(a) We advise to give a strong role to existing ESFRIS for domain specific training and capacity building and a strong role for the existing and collaborating e-INFRA groups for more generic ICT related aspects.
(b) We advise to develop a policy for 'certification' of institutes and consortia in and across member states that will be 'recognised authorities with the mandate to 'sign off' on DMPs
(c) We advise to ensure that there is a clear separation between the 'entity' that 'signs off' on the organisation that 'executes' the DMP (and will thus receive funds from the awarded project grant').
(d) We do not preclude that this could be the same organisation (or in the future a dedicated group in the research institutions involved), but it should not be an 'automatism' that the co-designer and 'endorser' of the DMP will also execute it. In other words, the PI's of the project will have optimal freedom to choose the 'best services around' to execute the DMP, including private services that meet the Rules of Engagement of the EOSC.

Recommendation

We urge the European Commission to address this particular action with the first level of urgency to make the EOSC a reality. Investments (via the challenge scheme) in this type of capacity in the member states will raise awareness of the enormous importance of good data stewardship for discovery and will (separate from dedicated de novo investment in infrastructure) mobilise billions of EUROS for proper data stewardship from existing research funds, including H2020. It will also propel Europe to the forefront of dat driven science by building the crucial expertise in the MS without additional burden on researchers and review panels and processes.

-