EOSCsecretariat.eu
Setup and management of the EOSC Secretariat supporting the EOSC Governance

Report on the online session on visions, requirements and needs for Future Research Environments in the Healthcare domain

Iain Buchan (University of Liverpool), David Carr (Wellcome Trust), Katharina Flicker (TU Wien), Sir Peter Ratcliffe (The Francis Crick Institute, Oxford University), Andreas Rauber (TU Wien)

1. Introduction

The European Open Science Cloud (EOSC)[1] initiative aims at supporting more than 1.7 million researchers and fostering interdisciplinary research in Europe. To understand better, what the research community needs, the EOSCSecretariat[2] partner TU Wien[3] is organizing workshops[4], interviews and consultations. As part of this initiative, the EOSCSecretariat invited a group of three top-level researchers and research enablers from the Healthcare domain to online sessions in June 2020 on visions, requirements and needs to inquire not only about their actual needs with respect to research infrastructures, services and policies, but also about visions and future needs and requirements. The objective was to:

- bring together a small, focused group of excellent researchers and research enablers who have a vision for and are interested in shaping the future of European research infrastructures for their domain
- identify current barriers and services considered essential for a well-functioning EOSC
- obtain a better understanding on how research is changing
- elaborate visions on how research will be conducted in 5 to 15 years and what the effect and impact on research infrastructures will be
- provide these as seeds for public comments to involve a large stakeholder community, thus ensuring many voices are being heard and that findings are considered in EOSC implementation processes, so that the EOSC is serving our needs as researchers

All findings of the discussion are shared with potential stakeholders and feed directly into the work of the EOSC governance bodies and the EOSC Working Groups (WGs), providing input crucial for the development of the EOSC.

2. A “wish list”: Identified Key Services

2.1 Reshuffling Research Environments

2.1.1 Data Literacy (EOSC@School)

The undervaluation of education on a societal scale is a challenge as it also puts research at risk. It limits the possibilities of science, and makes it difficult - or even impossible - to reach the frontiers of knowledge and go beyond. Research should not be driven by fashion and hypes: topics that are not prominent now may turn out to be essential sometime in the near future. We need mechanisms to ensure that a certain breath of research

[1] See https://ec.europa.eu/digital-single-market/en/european-open-science-cloud

[2] See https://www.eoscsecretariat.eu/

[3] See https://www.tuwien.at/en/

[4] For more information on the workshops, see https://www.eoscsecretariat.eu/co-creating-eosc-workshops-researchers-and-university-networks. Full reports on the workshops can be downloaded at https://zenodo.org/record/3693914#.XwhIhSgzaUk (workshop with university networks) and at https://zenodo.org/record/3701194#.XwhI0CgzaUk (workshop with researchers with a focus on technical and natural sciences). For Key-takeaway messages on these workshops and a discussion with funding bodies, please see https://zenodo.org/record/3701269#.XwhIyygzaUk.

activities is preserved. Thus, EOSC services should not be limited to supporting mainstream-topics, but also address the need of topics that are – at the moment – considered niche.

In addition, the need to not only teach scientists, but also pupils and students basics in data science is gaining importance as people need to have competences in reading data, evaluating data, classify information correctly, or assess whether data is suitable for addressing a specific task. Thus, **data literacy basics should be anchored in curricula early on and form an integral part of EOSC training at all levels of education**.

2.1.2 Support for the whole research cycle

The whole research cycle is in need of support. This includes the services mentioned in 2.3.1 Trustworthiness of data. In order to do cutting-edge research, scientists need to be able to trust the quality in the data they use. Apart from that, **evaluation services to check the quality of data, research output or verify experiments and results** are crucial to have.

In addition, AI and machines may be of use when it comes to peer-review processes. For example, AI approaches have value in **supplementing and reinforcing human peer review** in terms of e.g. running data, software checks, checking biases, reporting standards and potentially some elements of reproducibility. Thus, they could assist in quality control, give access to primary data and support trust in the conclusions derived. In the more distant future, it is crucial to ask if **machines could help to overcome human biases and thus would be better at peer reviewing** as algorithms can be inspected and evaluated, offering – in principle – higher transparency than human decision making processes.

However, it should be noted that even though AI is a powerful tool, it might be dangerous to rely too much on it (e.g. risk of losing scientific methods). Thus, solutions for e.g. in-built quality controls, and the exact use of AI and machines as well as mechanisms and processes to inspect, double-question and verify automated decision-making processes have to be provided.

2.1.3 Bigger is not necessarily better

Big data offers great statistical power. However, getting the bigger dataset and drawing conclusions from that does not necessarily enhance science or scientific output, because big data also comes with higher complexity. Higher complexity in return might lead to a higher false discovery rate. Thus, while having access to big data is useful it is also crucial to have **support in EOSC to conduct highly focused studies**.

2.1.4 Scientific Expert Programs

Having **support teams for scientific experts** is helpful and would speed up scientific processes. Researchers could then - depending on the expertise of such teams - contact experts in e.g. statistics, programming, data stewardship, and many more, if questions and requirements outside their core domain expertise should arise. These support teams do not necessarily have to be localized within the institution of the researcher but could form part of a **"human expert infrastructure" service** of the EOSC. However, it is important to note that when extensive input is needed, it makes sense to have the skills embedded in the team.

2.1.5 Enhancing collaboration

Research collaborations and research consortia facilitate pooling resources in order to achieve common goals. However, there are limits to this kind of joint efforts: involving and synchronizing all key players efficiently is

EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, grant Agreement number 831644

just not possible, if consortia are too big. Thus, **research collaborations within smaller groups must be facilitated and funded**.

2.2 Facilitating Data Intensive (Health) Systems

2.2.1 Connecting Health Systems

Covid-19 brought to extreme focus, that many systems do not have the intelligence to take timely actions to control or contain a pandemic, or to **reconfigure their systems to allow cross-agency communication** for agencies to talk and **support concepts of actions across different agencies**, which is crucial for fundamental public health. The data and all the basic technologies to respond better exist (e.g. crowdsourcing, peer-sourcing, basic analytic methods, augmentation or data-collections). Nevertheless, only very few high-level research groups study combinations of mechanisms and combinations of disease outcomes and inputs which results in a lack of understanding of real-world health systems.

Connecting various health systems with each other is key in order to understand such systems better because it allows bringing in different communities and considering socio-economic deprivation. This is difficult to achieve not because it is technically hard, but because it **requires organized efforts of society**.

National trusted nodes, however, are a good starting point to connect different health systems by working on specific topics, aligning with each other, flexibly agreeing on e.g. protocols on the technical, legal and trust level and – by doing so – enabling an **organic growth of research institutions, health provider networks**.

This kind of organic growth may also help to **reduce dangers that come with huge (European) projects**, such as the ever-seeking bigger numbers in disease specific registries (which is easy to do and not where the big unknowns are), or confusing people with discussions of objectives that are abstract and far in the future, instead of focusing on the smaller steps that have to be taken along the journey and that can be explained together with their societal benefits.

2.2.2 Counteracting Data Colonialism

In a connected world, with universal sharing/data access being a possibility, it becomes far too easy for nodes with many resources to exploit the data of another node, and to capitalize the information that comes with it. Thus, **counteracting data colonialism** is a political imperative. Potential solutions must involve society and need to be based on **trust and mutual aid**. In addition, solutions need to be small enough to gain the population's support on e.g. advanced data sharing, but big enough to have some statistical power.

There are many ways to gain the population's trust. For one thing, it is crucial to **have the data work for residents** and do better for society with what is already there. Implementing consortia of actions and protocols, impeding vendor dominance, establishing basic GDPR rights to data portability, duties of transparency and cooperation for each participant, or relying on civic data cooperatives[5] further detailed below are other ways of building up trust and increase access to health care data with an active consent of citizens in the long term.

[5] The Liverpool Civic Data Cooperative (CDC), for example, is a SIF-funded project. For further information on the project and CDCs, please see https://www.liverpoolcityregion-ca.gov.uk/liverpool-city-region-combined-authority-announces-proposals-for-5-3m-funding-for-data-driven-health-improvements/, https://metrolabnetwork.org/liverpool-hosts-international-symposium-building-smart-cities-with-citizens-and-for-the-public-good/, https://www.liverpool.ac.uk/population-health-sciences/news/articles/report-outlines-strategy-to-meet-government-mission-of-five-extra-healthy-years-by-2035 and https://documentcloud.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3Aa0271ed8-f813-4bb3-a0e4-342ab024e58f#pageNum=3.

Building **consortia of actions and protocols** supports the scaling and **managing of actions as well as steering discussions that are focused on results**, rather than on fear and insecurity. The latter would very likely result from building consortia of data or a database as people tend to get very protective of "their" data. Thus, conversations could not serve societal benefit as they would get prematurely stuck.

2.2.3 Civic Data Cooperatives

Trusted health systems that come along with societal, transparent benefits, are more likely to be supported by citizens (see figure 1). The current situation, however, is one of disease specific registries with highly restrictive access to data, where powerful and data-protective researchers effectively form cartels. They benefit from restrictiveness, because actual reward systems put great emphasis on e.g. first publications instead of stressing benefit-increasing mechanisms such as maximizing knowledge or rewarding researchers for the utility of information used from the data, they are supposed to be curating.

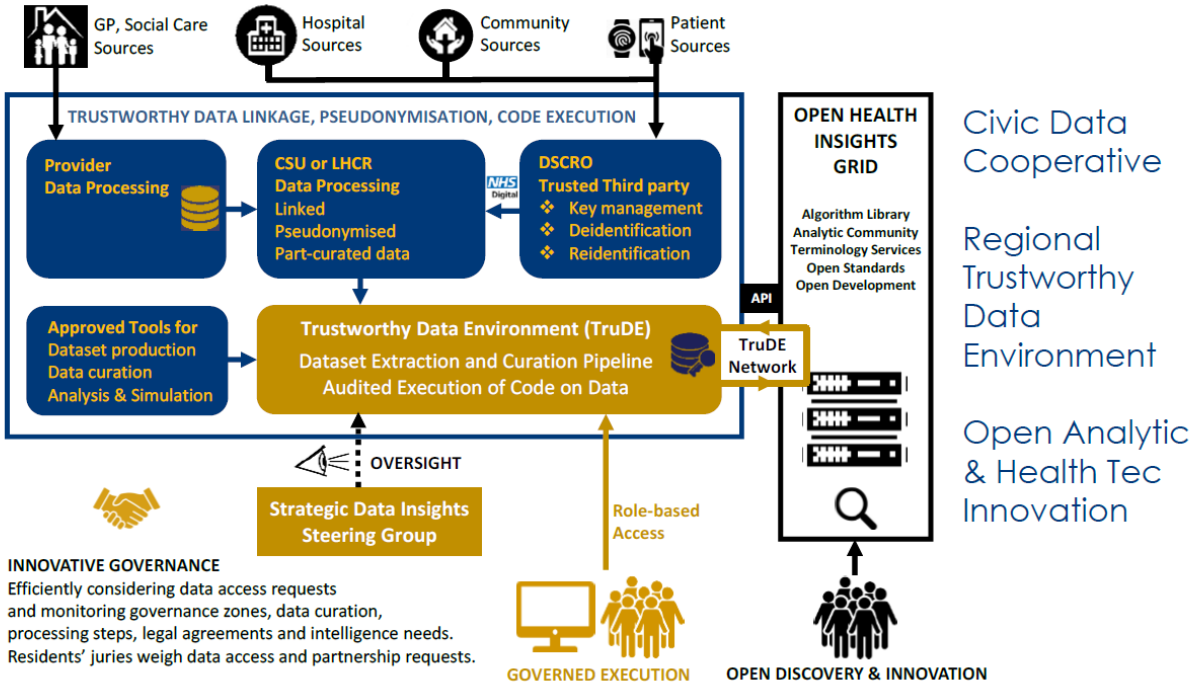

Figure 1: A common trustworthy data environment (Source: Buchan, Iain. 2020: The Health of the Nation. A strategy for Healthier Longer Lives. https://documentcloud.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3Aa0271ed8-f813-4bb3-a0e4-342ab024e58f#pageNum=1).

Thus, **citizens need to act as their own hub and as a trusted third party** without the incentive of forming cartels in order to **optimize data uses** (see figure 2), which includes ensuring the data flow where they are not flowing because of the aforementioned data cartels. This is key because in order to derive increasing societal benefits from the use of data, data has to be taken away from those with an immediate (economic) interest in data (such as research institutions and health systems) and put into a neutral trusted body that only benefits from data being used. Such **civic data cooperatives** are not about radical changes but about **developing a culture that starts to think more as a group or as a system**. They are, in a way, like a power station on a grid: Interoperability standards and shared infrastructure are the power lines and pylons. The whole grid is greater than its sum – there is no hub power station for data – just a cooperative way of working between nodes and basic, consistent, persistent shared infrastructure.

EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, grant Agreement number 831644

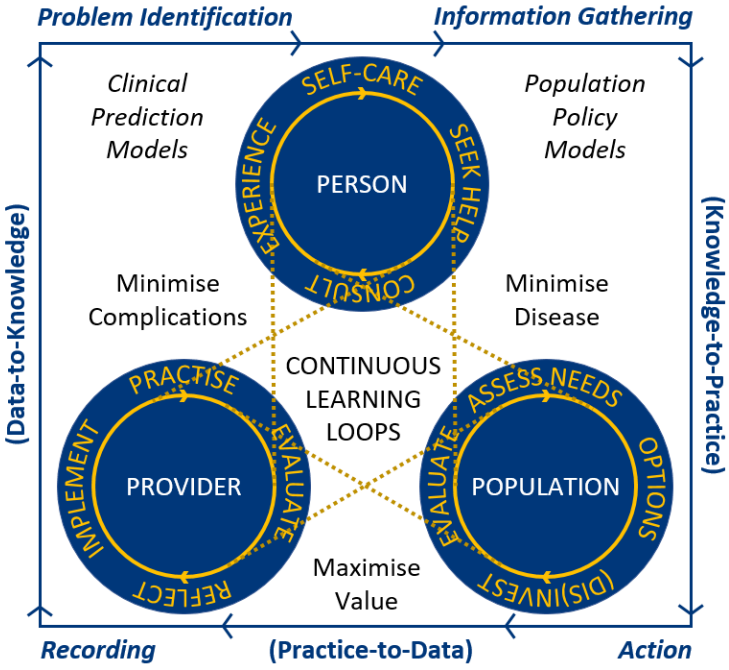

Figure 2: A Social-Value Data Engine. Optimizing the use of data through sharing and enriching system-wide data that involves all communities in a social-value data engine relies on support by citizens (Source: Buchan, Iain. 2020: The Health of the Nation. A strategy for Healthier Longer Lives. https://documentcloud.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3Aa0271ed8-f813-4bb3-a0e4-342ab024e58f#pageNum=1).

2.2.4 Avoiding vendor lock-in

Data intensive systems must be able to interact. Thus, it is crucial to **avoid vendor lock-in and vendor dominance.** The former requires strict enforcement of **interoperability**. For one thing, **tool chaining** that allows for flexible processing workflows, or **interoperability sessions** counteract vendor lock-in. Interoperability sessions can be scheduled on a regular basis (e.g. biennially) for three to five working days and oblige computing vendors to get together to demonstrate tool chaining the lossless exchange of data and prove their interoperability. This kind of **approach to bottom-up integration** could be established for all kinds of services.

2.2.5 Avoiding vendor dominance

The **growth from a bottom-up network** in combination with basic **GDPR rights to data portability** and the **duty of transparency and cooperation for each participant** support the interaction of systems. If common basic principles and frameworks allow a local innovation group to self-organize and extend an intelligent health system without requiring the permission of the major vendor, or swap services very quickly without being locked in, then **systems will evolve and become resilient** to vendor dominance. Similar to the avoidance of data colonialism, dominance of vendors or institutions providing specific services need to be avoided. This may require controls on market share to avoid quasi-monopolies emerging for certain services.

EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, grant Agreement number 831644

2.3 Data

2.3.1 Trustworthiness of data

The trustworthiness of data has to be ensured by all means. However, trust requires understanding where data comes from. Consequently, there needs to be a transparent **framework to understand the data**. That framework would have to depend on the type of data. In general, provenance and data quality has to be checked or, at the very least, be checkable.

Thus, EOSC will require **services to help with checking and verifying data quality**. For, example, metadata and provenance tracking are crucial. Researchers need to be able to assess the quality of data that they consider to use. In addition, the freedom of use of data is highly important, though not all data can (and should) be opened as it may be sensitive data that cannot be anonymized. Hence, the following services[6] may come in handy:

- Services to enable the automatic recording of provenance metadata for data, computation, and processes
- Services for data capturing
- Services to track provenance
- Services for secured and monitored data visiting
- Services for validation processes

A short description of these services can also be found in chapter *3. Table: Collection of Identified Services* further below.

2.3.2 Reproducibility in Research: Access to primary data

Articles on the lack of reproducibility of research results, as well as recommendations on how to make research reproducible were already published years ago[7]. However, the reproducibility of research is still an issue. EOSC **needs to support robust research by providing access to primary data**, which is crucial in reproducing results of scientific research, and mechanisms to support reproducibility studies.

2.4 Interdisciplinary Research

2.4.1 Translation Services

Given the complexity of today's world and problems that cross borders, cultural divides, the many different fields of knowledge and the need to somehow bring them all together, the future of research will definitely be strongly interdisciplinary. However, **translation services to help communication across discipline boundaries are needed** in order to do interdisciplinary research efficiently. Central key terms and scientific concepts have to be translated from one discipline to another. In addition, these key terms, scientific concepts as well as research outputs have to be communicated to policy makers and to the public.

[6] All of these services have also been identified at a workshop on future research environments that was organized by TU Wien and held in January 2020. The full report including a list of all collected services can be found here: https://zenodo.org/record/3701194#.Xu0fHWgzaUk (DOI 10.5281/zenodo. 3701194).

[7] Examples are the article *1.500 scientists lift the lid on reproducibility. Survey sheds light on the 'crisis' rocking research* published in *Nature* in 2016 (https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970) and *Ten simple rules for reproducible computational research* published in PLOS in 2013 (https://doi.org/10.1371/journal.pcbi.1003285).

EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, grant Agreement number 831644

The need for these kind of **translation services** has already been identified at a TU Wien-organized workshop on future research environments in January 2020[8]. Back then, these services were labelled **services for horizontal translation & communication adaptation**. A more detailed description of these services can be found below in section *3. Table: Collection of Identified Services*.

2.4.2 Knowledge Brokering

In order to recognize and use opportunities for research to be translated into tangible benefits for society, **services to facilitate two-way or multiway exchange of information** are key. Pieces of knowledge have to be brought together in order to bridge gaps between knowledge producers and knowledge users (e.g. scientists and people working in the health domain) as well as to create knowledge infrastructures.

All of that closely relates to the need of the aforementioned translation services because research outputs have to be communicated to different audiences. In addition, established and reliable **knowledge infrastructures will help to coordinate research efforts so that scientists look at different aspects of one problem** at the same time, rather than having a look at the same, most prominent aspect. In return, a broader perspective on a scientific problem will provide a more comprehensive understanding of a problem and thus likely result in better solutions.

Knowledge brokering also matters because it helps not only with the communication of research outputs to different audiences, but also promotes them. Thus, it supports increasing research impact in society. In a way, services for knowledge brokering can be associated with **services for research promotion**[9] that emphasize the need of EOSC moving beyond data, or data infrastructures and recognize its importance of laying the foundations for the future production of knowledge and knowledge transfer.

[8] A full report on this workshop and its results is available at https://zenodo.org/record/3701194#.Xu0fHWgzaUk (DOI 10.5281/zenodo. 3701194).

[9] **Services for research promotion to increase the impact of science in society** are services that have been identified as requirements at a workshop called "What do our researchers want? A workshop to "collect" needs, requirements and visions on ideal future research environments". It was held in January 2020 and organized by TU Wien. For further information, please see the full report: https://zenodo.org/record/3701194#.Xu0fHWgzaUk (DOI 10.5281/zenodo. 3701194).

EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, grant Agreement number 831644

3 Table: Collection of Identified Services

Services	Description
EOSC@School / Data Literacy	People need to have competences in reading data, evaluating data, classify information correctly, or assess whether data is suitable for addressing a specific task. Thus, data literacy basics should be anchored in curricula early on and form an integral part of EOSC training at all levels of education.
Services to support human peer review processes	AI approaches have value in supplementing and reinforcing human peer review in terms of e.g. running data/software checks, checking biases/reporting standards and potentially some elements of reproducibility. Thus, they could assist in quality control, give access to primary data and support trust in the conclusions derived. Eventually, it is crucial to ask if machines could help to overcome human biases and thus would be better at peer reviewing as algorithms can be inspected and evaluated, offering – in principle – higher transparency than human decision making processes.
Incentives for and recognition of highly focused studies	Drawing conclusions from big data does not necessarily enhance science or scientific output, because big data also comes with higher complexity. That might lead to a higher false discovery rate. Thus, while having access to big data is useful it is also crucial to have support in EOSC to conduct highly focused studies.
"Human expert infrastructure" services	Scientific expert programs support researchers, who can contact experts in e.g. statistics, programming, data stewardship, and many more, if questions and requirements outside their core domain expertise should arise.
Services and funding mechanisms to facilitate research collaborations within small(er) groups	Research collaborations and research consortia facilitate pooling resources in order to achieve common goals, but there are limits to this kind of joint efforts: involving and synchronizing all key players efficiently is not possible, if consortia are too big. Thus, research collaborations within smaller groups must be facilitated and funded.
Services and organizational incentives to connect various health systems	Connecting various health systems with each other is key in order to understand such systems better because it allows bringing in different communities and considering socio-economic deprivation. This is difficult to achieve because it requires organized efforts of society. National trusted nodes, however, are a good starting point to connect different health systems by working on specific topics, aligning with each other, flexibly agreeing on e.g. protocols on the technical, legal and trust level and – by doing so – enabling an organic growth of research institutions, health provider networks.
Rules and services to counteract data colonialism	Counteracting data colonialism is a political imperative. Potential solutions must involve society and need to be based on trust and mutual aid. Thus, it is crucial to have the data work for residents and do better for society with what is already there. Implementing consortia of actions and protocols, impeding vendor dominance, establishing basic GDPR rights to data portability, duties of transparency and cooperation for each participant, or relying on civic data cooperatives are ways of building up trust and increase access to health care data with an active consent of citizens in the long term.
Initiatives to optimize data uses that are supported by citizens	In order to derive increasing societal benefits from the use of data, data has to be taken away from those with an immediate (economic) interest in data (such as research institutions and health systems) and put into a neutral trusted body that only benefits from data being used. Thus, citizens are needed to act as their own hub and as a trusted third party. Such civic data cooperatives are not about radical changes but about developing a culture that starts to think more as a group or as a system.
Rules, protocols and verification mechanisms to avoid vendor lock-in	It is crucial to avoid vendor lock-in and vendor dominance, because data intensive systems must be able to interact. Thus, interoperability must be enforced: Tool chaining that allows for flexible processing workflows, or

EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, grant Agreement number 831644

	interoperability sessions can counteract vendor lock-in. Such Interoperability sessions can be scheduled on a regular basis (e.g. biennially) for three to five working days and oblige computing vendors to get together to demonstrate tool chaining the lossless exchange of data and prove their interoperability. The growth from a bottom-up network in combination with basic GDPR rights to data portability and the duty of transparency and cooperation for each participant support the interaction of systems.
Rules to avoid quasi monopolies and dominance	Control of market share and ensuring that no contributor of services becomes dominant in size or too-big-too-fail is crucial. In combination with mechanisms countering lock-in into specific services or providers, the EOSC should ensure resilience and flexibility by guaranteeing that a diverse set of actors are providing services in a transparent and cooperative manner.
Services to help with checking and verifying data quality	The trustworthiness of data has to be ensured by all means. However, trust requires understanding where data comes from. Consequently, there needs to be a transparent framework to understand the data. In general, provenance and data quality, has to be checked or, at the very least, checkable. EOSC shall offer services that allow for the documentation of how data is gathered. It should make (meta) data more searchable, findable and trackable and it should support the many different formats. Examples are services to enable the automatic recording of provenance metadata for data, computation, and processes, services for data capturing (that can even include cryptographic verification of data like cryptographic signatures on the metadata on the data being processed), and services for trusted validation processes[10]. The latter is crucial because researchers will refrain from using datasets that they do not trust. Thus, EOSC must find ways to guarantee the quality of data.
Services for secured and monitored data visiting	Such services need to support the negotiation of access to data automatically. Trust analysis algorithms need to be established and leakage of information needs to be prevented.
Services to support robust research by providing access to primary data	Access to primary data is crucial in order to reproduce research results. Thus, EOSC should provide researchers with the means to access primary data and provide services that support reproducibility.
Services for horizontal translation & communication adaptation	Translation services to communicate research outputs to policy makers, to the public and across disciplines, e.g. services for automatic metaphor translation. Services to translate scientific concepts and explanations for different levels of expertise within a discipline, e.g. services for math translation explaining some mathematical concepts for specific research questions.
Services for knowledge brokering	Such services facilitate two-way or multiway exchange of information, bridge gaps between knowledge producers and knowledge users and build connections between different audiences. Thus, they also promote research output, increase research impact in society and steer research towards a more comprehensive understanding of a problem.
Services for research promotion to increase the impact of science in society	EOSC is not only about data, or data infrastructures. It is also about laying the foundations for the production of knowledge and knowledge transfer (teaching, mentoring, and educating). Thus, services to promote science and increase its impact in society are needed.

[10] A more detailed description of these services can be found here: https://zenodo.org/record/3701194#.Xu0fHWgzaUk (DOI 10.5281/zenodo.3701194)