Report of the European Commission Public Consultation on Open Research Data

Executive summary and key outcomes

The European Commission held a public consultation on open research data on 2 July 2013 in Brussels, which was attended by a variety of stakeholders from the research community, industry, funders, libraries, publishers, infrastructure developers and others. The debate focused on five questions posed by the Commission to structure the debate and can be summarized as follows. Information on the consultation, including the agenda, the list of participants, the list of contributions and the final report are available here: http://ec.europa.eu/digital-agenda/node/67533.

1) How can we define research data and what types of data should be open?
Definitions of research data vary, with some contributions defining research data as potentially all data (including public sector information), and some limiting it to data that is the product of research. From the perspective of researchers, research data includes all data from an experiment, study or measurement, including metadata and details on processing data. For publishers, data linked to publications is part of the publication.

2) When and how does openness need to be limited?
Potential limitations are connected with issues of public security, privacy and data protection, as well as intellectual property right (IPR) protection and possible commercialisation. Concerning public security, the potential use of data for terrorism was mentioned. Privacy and data protection are seen as particularly relevant in areas like health, in particular for clinical trials and the issue of opening up negative results. For IPR and possible commercialisation of research results, representatives from industry expressed the view that data resulting from projects that are close to market should as a rule not be open, but may be opened on an individual case-by-case basis.

3) How should the issue of data re-use be addressed?
This question led to discussions about licensing, but also about technical aspects of open research data. The discussion centred not just on whether and how data should be re-used, but also on the adequacy of e-infrastructures for data re-use. In the context of re-use, the Directive on the re-use of public sector information (2003/98/EC, currently under revision) was mentioned several times. While public sector information (PSI) is distinct from research data and governed by a specific directive, it is important to remember that this type of information can also be useful for research.

4) Where should research data be stored and made accessible?
The need for improved data management practices and better data accessibility is a key concern. These issues are closely linked with data preservation and the sustainability of data repositories for data. The readiness of professionals to engage in data curation was also highlighted. All stakeholders agreed that any funding body policy on open research data must call on researchers to take the issue of data management seriously by developing data management plans (DMPs) for their research projects.

5) How can we enhance data awareness and a culture of sharing?
Stakeholders consider data awareness and a culture of sharing to be one of the most important - if not *the* most important – but probably also the most difficult to address in formulating policy on open access to research data. One important way in which data awareness and a culture of sharing can be spread is by establishing mechanisms and processes to recognise and reward and even require good data sharing practices. Making it possible to publish and cite data (for example in data journals) is one major way forward in this respect.

Report of the Consultation

Background

Technology is quickly changing the scope and potential of research, and this requires new policies in specific areas. One of these concerns the area of research data and the possibility of opening up research data in the digital age. In order to hear the views of stakeholders, the Commission (DG CONNECT and DG RTD) organised a public consultation on open research data, consisting of (i) a one day event where individual presentations and discussion could be heard (held on 2 July 2013 in Brussels), and (ii) a written consultation period (19 June to 15 July 2013). Some 130 participants participated in the event 2 July event and 45 contributions were received.

Five lead questions were posed by the Commission:

1) How can we define research data and what types of data should be open?
2) How should the issue of data re-use be addressed?
3) When and how does openness need to be limited?
4) Where should research data be stored and made accessible?
5) How can we enhance data awareness and a culture of sharing?

The following stakeholder groups were represented during the 2 July event and via written contributions: researchers, industry, research funders, libraries, publishers, infrastructure developers and a separate category for "other voices" in order to facilitate contributors who fell into none of the above. The Commission was able to accommodate all requests for participation.

While the discussion was broken down into individual stakeholder groups, these minutes follow the five questions posed. This format is intended to allow for the individual interests to be balanced clearly and effectively. These minutes will be published on the Commission website along with the slides presented and the written submissions received.

Welcome and introduction by the European Commission

The event was opened by Thierry Van der Pyl, Director of Excellence in Science at DG CONNECT and Octavi Quintana, Director of the European Research Area, DG RTD. Thierry Van der Pyl provided an overall framework for the event, stating that the Commission strongly advocates open science policies. He said that a policy on open research data can benefit science, business and citizens. Octavi Quintana introduced the five questions, which had been disseminated before to structure the event and the contributions.

1) How can we define research data and what types of research data should be open?

This question was addressed by all groups of contributors. Definitions of research data varied, with some contributions defining research data as potentially all data, and others limiting it to data that is the product of research and/or data that is used for research. From the perspective of researchers, research data are all data from an experiment, study or measurement, including the metadata and processing details. There was some acknowledgement that data should be restricted in certain cases, such as sensitive data and commercial data.

Researchers stressed the importance of data management plans (DMPs) and took the viewpoint that all data should be open by default. The Netherlands Organisation for Applied Scientific Research (**TNO, Department of Microbiology and Systems Biology)** noted that we need to consider research data resulting from participatory research and metadata. The **Istituto Superiore di Sanità** made the point that while research data and open data are closely connected, the different nature of types of research data needs careful consideration, and also raised the point that data awareness needs to be created. The **Erasmus Universiteit Rotterdam** also mentioned the danger of making some data open, referring to the specific example of genome sequencing. The question of patents was also raised during this contribution. **CERN** stressed the need for developing incentive systems and opportunities that would allow researchers who have data to share to be able to do so and be recognised. The **European Consortium of Innovative Universities** referred to differing needs and traditions in different fields.

Industry views research data as being potentially any type of data. The viewpoint put forward by **Philips** is that research data should be open on a case-by-case basis, particularly for research that is partially funded by industry. The **Federation of German Security and Defence Industries** put forwards four categories of research data related to research projects: biographic data of authors, metadata of the research project, the executive summary and the full text of the final report. While the former three may be open, the latter as a rule should not be opened except on a case-by-case decision by the owning industry partner.

For **Research Funders**, data is considered a public good, and it must be possible to find and search it (as expressed by **Research Councils UK)**. The **Wellcome Trust** said that it would like both publications and data to be openly available. The objective is to maximise benefit to the public. The **Netherlands Organisation for Scientific Research (NWO)** stated that narrowing the concept of data to any specific type of data would not work and agreed with a broad definition of research data. The **Open Knowledge Foundation** state in their written submission that research data is extremely heterogeneous and that it takes a variety of forms including numerical data, textual records, images, audio and visual data as well as custom written software, other code underlying the research and pre-analysis plans. They also argue that it should include metadata.

The **Information Systems & e-Infrastructure** perspective also contributed a range of views. The **Institute for the Study of Labour** proposed that research data is any data, and compared this to clues at a crime scene. They offer two definitions for research data, which are either "data suitable for research" or "any data that has been used to answer a research question is research data". The infrastructure project **OpenAIRE** added that when you see a piece of data, it is not necessarily clear whether it is research data or not. The **Data Archiving and Networked Services (DANS)** said that it is important not to focus only on big data, but also small data and in many different disciplines. It also stated that disciplinary priorities need to be taken into account, suggesting the principle "open if possible, protected if necessary". The **Joint Information Systems Committee (JISC) and Knowledge Exchange** underlined the importance of DMPs. In their written submission, they state that research data exists in many forms and in various discipline-specific varieties of formats and data types. They elaborate to state that different types of research take different forms throughout the research process, and that it may not be realistic or effective to provide open access to all of these data sets. The project **pro-iBiosphere** defines research data as "a collection of logically connected facts (observations, descriptions or measurements), typically structured in tabular form as a set of records, with each record comprising a set of elements and recorded in one or more computer data files along with metadata that together comprise a data package". They argue that the vast majority of biodiversity data should be open, and draw attention to what they refer to as the "often neglected" data in PDFs and paper publications.

Publishers took varying views on what data is and how it should be open. **eLife** suggested that publicly funded research data is a public good and that it should be shared effectively to maximise the impact of public research funding. In their submission, they noted that "research data are primary outputs of a research process that are intended to be incorporated into research communications as support for the claims of that research" and suggested that all data should be open as a default, that CC0 and CC-BY licences can be useful to share data, and that data must be accessible as well as legally and technically usable. **Elsevier** contributed to this debate stating that we need to distinguish between the digitisation and curation of research data on the one hand, and the process of sharing it on the other. The point was also made that while some data cannot be open, due to "privacy issues, human-subject studies and patent and intellectual property claims", that does not mean that users should not know it exists. The use of data catalogues was proposed as part of the written submission from **Elsevier**.

Libraries subscribed to the approach that research data should be open by default. The **European University Institute** referred to very strong disciplinary, and sub-disciplinary differences between research data cultures. The **TU Delft Library** suggested that data enriching activities should be included for funding under Horizon 2020. **EBLIDA** suggested that data can take many forms, ranging from pure data, such as scientific information created in a laboratory or clinical trials, to unique data found in journals, newspapers, books, and the web.

Other Voices In their submission, **Creative Commons** takes the view that all data required to reproduce or verify a published research result should be made open and where such data are a part of a bigger data set, the entire data set should be made open. The **European Public Health Alliance (EPHA)** state that they support a broad definition of research data. **EPHA** disagrees with concerns raised on the loss of competitiveness that could result from hurt commercial interests that may result from open access and open data. The European Commission took this into consideration in the impact assessment of the ERA, which showed that open access and open date would have a positive benefit on the economy and competitiveness. The **International Council for Open Research and Open Education (ICORE)** states that as much research data as possible should be open, and that all publicly funded research data should be open.

2) When and how does openness need to be limited?

This issue came up throughout the day. The limitations mentioned were connected with public security, privacy, intellectual property rights and commercialisation. Privacy is relevant particularly in areas like health where the results of clinical trials and negative results were mentioned a number of times. For public security, organic chemistry was mentioned with regard to concerns over the use of data in terrorism while genome sequencing could raise ethical issues such as genetic engineering. Intellectual property rights are another central issue, in particular in relation to the concerns of industry, publishers and the research community over the varying roles of intellectual property rights. For example, representatives from industry expressed the view that the ownership of patents raises issues over the ownership of data (despite data often referring to facts which cannot be copyrighted). For publishers, the data may underlie the publication and therefore be part of the publication and thus part of the expression of those facts. Finally, for the research community the role of moral rights and control over how research data is used raises the most concerns.

Researchers. **TNO** stated that unpublished and commercial data *may* be closed, and that sensitive data *should* be closed. The **Erasmus Universiteit Amsterdam** raised the issue of genome sequencing, which represents complete personal profiles. The **Helmholtz Association** referred to the need for balance between the privacy of data and the accessibility of research. **EMBL-EBI/ELIXIR** said that the term open data may be misleading and offered the term 'accessible data'.

Industry's viewpoint is that the availability of data should be assessed on a case-by-case basis. There was reference to the view that data is different from publications and should not be open by default. In the discussion, the point of public funding was raised and the role of openness in public/private partnerships. It was suggested by industry representatives, for example **Philips**, that even when research is partially publicly funded, it should not be made available by default, as the research necessarily is also partially funded by the industry, who is dependent on the commercialisation of the output.

Research Funders: **Research Councils UK** referred to the importance of "discoverable" data, but also said that discoverable does not mean that anyone can access data. The **Wellcome Trust** pointed out that not all data has the same value, and that, from a funder perspective, value judgements must be made. It also referred to the limits of data sharing, such as the role of intellectual property rights, and made the point that different types of data would raise different issues. The **Swedish Research Council** made reference to public universities where data is considered public information, but also emphasised the need to balance with protecting personal integrity. In its written submission, **NWO** stresses that, in the case of research projects run as public/private partnerships, an agreement between publicly funded and private parties should be set up defining "where open access to the data needs to be limited and what licences will be in place". **NWO** further agrees with the statement that the default for research data should be open and notes that situations in which access to data needs to be limited have already been defined in previous European Commission open access policies, including privacy, public safety and intellectual property rights. The **Alliance of German Science Organisations** states in its written submission that the protection of personal data of those affected by collected data, as well as obligations to third parties must be taken into account. The **Open Knowledge Foundation** acknowledges that while the default position should be that data is open, there are "situations where the full data cannot be released", citing the example of privacy.

Information Systems & e-Infrastructures. **OpenAIRE** suggested that openness could be limited to "quality" data, and also mentioned statistical confidentiality and the issue of qualitative interviews. **DANS** said that data collected with public money should not be privately owned and added that protection of privacy is a factor, but that it should not be a dogma. However, **DANS** argued that public interests and privacy interests should be protected. An embargo of up to two years before making data public was suggested. **JISC and Knowledge Exchange** made the point that important factors preventing data from being open include privacy and commercial considerations and stated that DMPs should address these issues. They state that "the default should be for data to be open, reusable and clearly licensed as such". **pro-iBiosphere** also states that the default for research data should be open and mention that an initial period of non-disclosure until publication of the research is acceptable.

Publishers. **eLife** supported making data available on CC0 & CC-BY licences, but also said that it should be restricted in certain cases. **The International Association of Scientific, Technical and Medical Publishers (STM)** stated that data must be understandable, and it is important that there is a good connection between data and publications, but the biggest fear is that data could be misused. **Elsevier** mentioned the use of data catalogues to annotate data that cannot be made available. In the discussion **STM** made the distinction between factual data not being copyrightable and ownership of data.

Libraries. The **European University Institute** gave two possible reasons for not making data open. The first reason is data protection and relates to cases where data subjects are people, households or firms, but anonymised versions of datasets can be made available for public use. The second reason relates to limitations due to copyrighted databases. The **TU Delft Library** made the general statement that it is of utmost importance that science is given back to society. The **German National Library of Medicine** suggested that there should be a regulatory framework in which the scientist does not have to give up his distribution rights on the data in favour of the publisher when publishing a paper. It also suggested that research data in already printed material should be freely made available especially for data mining purposes.

Other Voices. In its written submission, **Creative Commons** takes the view that the default for research data should be open. Starting from that default, there may be a few reasons requiring that openness be limited including privacy and confidentiality of human subjects, cultural sensitivity and national security. **EPHA** referred to the existence of legal frameworks protecting commercial information. The view from **ICORE** was that there is no need to limit openness, but that instead efforts should go into structuring it through precise rules and regulations.

3) How should the issue of data re-use be addressed?

In response to this question, concerns about licensing and about recognition for researchers were raised. The question also led to more technical aspects of open research data as the discussion centred not just on how data should be re-used, but also on the adequacy and preparedness of e-infrastructures for data re-use. This question is therefore closely linked with question 4) on the storage and accessibility of data.

Researchers. **TNO** stated that people should be acknowledged if data is re-used and the **Istituto Superiore di Sanità** said that it is very important that researchers have incentives to share data and that funders should adopt binding mechanisms to ensure that data is open and re-used. The **Erasmus Universiteit Rotterdam** referred to data enrichment and the need to scrutinise the quality of data. **CERN** referred to the importance of a community of data sharing and suggested that "anyone who has data to share can do so and be recognised". **CERN** also praised the multi-stakeholder initiatives ORCID and DataCite in addressing the issue of recognition. The **Helmholtz** Association referred to the need for tailor-made solutions based on different disciplinary needs, but also openness and integrity. **EMBL-EBI/ELIXIR** stated that charging for data and seeking to restrict data would impede progress. The **European Consortium of Innovative Universities** also referred to the need for incentives for researchers. Regarding data re-use, four main areas to focus on were mentioned: tracking use, researcher integrity, data integrity and responsibility.

During the discussion, **LIBER** reiterated the need to convince researchers that their data is worth sharing and also the need to educate them about the basic elements. Peer review also came up in the discussion, with the **Alliance for Permanent Access** pointing out that there is too much data for traditional peer review to cope with. Crowdsourcing and the "Wikipedia approach" were cited as alternative approaches to peer review. **DANS** mentioned that big data should involve quality measures and archival practices. **eLife** mentioned that making data available gives the impression that the researchers did something and in terms of re-use referred to the level of support for the publication of negative results. **Wiley** stated that peer reviewers need to be clear on which part of the process they are involved in, with **DataCite /** the **British Library** stressing the importance of dedication and citation in peer review. **CERN** raised the point that scientific reputation is more powerful than peer review.

Industry. **Philips** reiterated that open access to research data should apply on a voluntary case-by-case basis. It was also said that forced open access to research data might hamper industry participation in Horizon 2020. The **Federation of German Security and Defence Industries** concurred with these views.

Research Funders. **Research Councils UK** stated that re-users have responsibilities in the use of research data. The **Wellcome Trust** added to this by stressing the need for balance between the needs of data generators and users, and stated the need to build extensive recognition mechanisms so researchers get a fuller picture of research outputs. Furthermore, the view was also expressed that funders should require researchers to maximise access to data of wider value, with as few restrictions as possible.

In the discussion, the **British Library** referred to the Directive on the re-use of public sector information (2003/98/EC, currently under revision) and pointed out that any intellectual property created by a university or library will be treated as public sector information. **ICORE** said there needed to be a clear distinction between the questions relating to copyright and the questions relating to licences. It also said that it would be helpful to set up standard patents. **Research Councils UK** added that terms and conditions should be set up from the start, such as the proposal time for the topic. **Philips** replied to this, stating that inventions protected by patents are on average worth more than those that are not. **Health Action**

International Europe contributed to the discussion by adding that all involved need to think about incentives for companies.

Several institutions and bodies also provided written submissions on this question. The **Finnish Ministry of Education and Culture** in their submission stated that "there should be jointly developed mechanisms and incentives for opening up data for re-use, whether data generated via research, or data which is valuable for research purposes". They also said that in order to improve re-use, there is a need to develop "common practices and operating models at the organisational level, co-ordinated nationally and in collaboration at the European and global level, focusing on interoperability and standards development". The **French National Centre for Scientific Research (DIST-CNRS)** observes that "the re-use of data will only become a working reality if datasets are perceived as trustworthy", adding that "it is essential that data creators define the property, use and citation rules to be applied". **NWO** stated that in the next Framework Programme, publicly funded projects should be "accessible, intelligible, assessable, reusable and referable, preferably by means of a persistent identifier". The need to develop standardisation regarding repositories was also identified. The **Alliance of German Science Organisations** states that the provision of research data for further use benefits the sciences and humanities in their entirety, adding that it encourages the recognition and support of this additional effort. The **Open Knowledge Foundation** notes that data is only "meaningfully open" when produced in a format and under an open licence that allows for re-use by other users. They argue that there is a role here for data publishers and repository managers in endeavouring to make the data usable and discoverable.

Information Systems and e-Infrastructures. One of the issues discussed by this group was that of Digital Object Identifiers (DOIs). The **Institute for the Study of Labor** stated that reusing data can both increase return on investment and safeguard scientific integrity, adding that the collection of all instances of the "same" experiment may help improve our understanding of how to connect them and how to do science in general. **DANS** suggests that data should be subject to peer review. They also suggested the creation of a persistent identifier. **JISC and Knowledge Exchange** argue that re-use for the purpose of science, education and business must be allowed. **FIZ Karlsruhe - Leibniz Institute for Information Infrastructure** suggested that publicly available data have to be easily findable and accessible, well-documented, understandable and quality-checked to ensure reliability. They also referred to the legal aspects, such as CC-licences or an equivalent licence for datasets.

Publishers. eLife emphasised once again that research data is a public good: "if data is to be made available with the intention of maximising the economic impact and the public good created, it is critical that re-use be enabled both technically and legally". It adds that data should be released under licences which maximize "the potential for re-use and recombination of that data. The appropriate licenses are the Creative Commons CC0 waiver and CC BY copyright licenses". **STM** believes that it is paramount that data and publications be closely inter-connected and integrated in order to ensure findability, accessibility, understandability and re-usability of research data. They argue that "data and publications must exist in a sustainable way ensuring longevity of open research data".

Libraries. The issue of re-use is discussed by the joint **LIBER, OpenAIRE and COAR**. Statement. They advocate that the use of "appropriate open data licences is highly recommended". Creative Commons CC0 is given as an example. They also state that research data underlying publications should be made available to the reviewers in the peer-review process. The **European University Institute** said that libraries have a "very important role to play in brokering open data issues between research teams and publishers". They said this can be achieved by using dataset metadata, which can help librarians determine archival requirements, levels of embargo and access. **The Max Planck Digital Library** suggested that continuous monitoring of re-use of datasets would decide if a set should be kept, raising the issue of the usefulness of the data to be stored, adding that continuous re-use of data is sign of scientific impact. The **TU Delft Library** states that "once the data are in the public domain, no royalties can be obtained for them nor can patents be obtained and therefore no exclusivity of exploitation rights can be granted". They elaborate to call for the "speeding up of the commercialisation process of the project IP ". They also propose that future publicly funded projects have "mandatory clauses according to IP dissemination and exploitation". **EBLIDA** make the point that "neither the United States or Japan in their statute based intellectual property laws protect databases of pure data, and through their respective limitations and exceptions regimes also allow the lawful extraction of data and facts that sit within text". The written contribution makes direct reference to the Licences for Europe exercise and the use of text and data mining technology to extract data from databases.

Other Voices. In their written submission, **Creative Commons** argued that "terms of use can be unambiguously conveyed by properly applying to the work a universally recognized public domain dedication mark" Examples given of this were CC0 or PDM, or a public license such as CC-BY, CC-BY-SA or ODbL. The web domain http://opendefinition.org/licenses/ was suggested for Conformant Recommended Licenses for “open data”. **ICORE** echoed this view, outlining that data re-use can be addressed "by providing simple legal licenses to select from (as templates)".

4) Where should research data be stored and made accessible?

This question was one of the most widely debated by contributors. The issue of DMPs and data accessibility was one of the main themes discussed. These issues are closely linked to data preservation. The sustainability of repositories for data and the readiness of professionals to engage in data curation were also points of discussion. There was a general acceptance from all stakeholders that DMPs are crucial for any policy on open research data.

Researchers. **TNO** asserted that data should be stored in a repository before disclosure and said that it is important that negative results also be included. The issue of a DMP was mentioned by the **Istituto Superiore di Sanità**, including the importance of establishing a global infrastructure. **CERN** emphasised the importance of listening to the research community needs of data-sharing and citing infrastructures. It advocated the need for an infrastructure that is both technical and social. As an example, **CERN** mentioned that researchers should be able to use repositories and drew the floor's attention to the DataCite & ORCID infrastructures. The **Helmholtz Association** referred to the need to develop tailor-made solutions and mentioned the need to include DMPs in applications for scientific funding. They also referred to the need for "safe havens" for the sharing of data and accessibility for research. **EMBL-EBI/ELIXIR** referred to the need for new policies since repositories already exist. The **European Consortium of Innovative Universities** made a further point regarding accessibility of data, stating that we need to be able to find data through a search engine and that we need to find ways of networking local databases. It referred to a good data centre as an infrastructure into which data can be entered and managed during a project and through which it can become open access at the push of a button. The **Academic Medical Center / University of Amsterdam** raised the point that publishing data in the public domain is expensive and suggested that stakeholders are the key enablers for this.

Industry. As has been mentioned above, industry representatives advocated a case-by-case approach regarding open research data, also regarding storage and accessibility. The **Federation of German Security and Defence Industries** identified three different levels of access to research data: 1) Basic (Fundamental Research): industry is involved on a case-by-case basis and participates more as an observer. Industry benefits from an open and unlimited access to all research data as it needs to obtain the results for its own strategic evaluation. 2) Applied Research: This is the first step towards a product approach and industry is strongly involved in these cases. Access to this information is decided on a case-by-case basis and full ownership of the information has to be guaranteed to industry. 3) Product Analysis Investigation: This is characterised by the exclusive involvement of industry and releasing these results to the public has to be denied in order to allow for commercialisation of the results by the industry.

Research Funders. From the perspective of research funders, **Research Councils UK** put forward the suggestion that data should more than anything else be "discoverable". The **Wellcome Trust** referred to the importance of DMPs and balancing the needs between data generators and users. The **Swedish Research Council** explained that in Sweden universities are responsible for archiving data from scientists, but this places a heavy burden on universities and the ideal situation would be to provide infrastructures that are cost efficient and heterogeneous. In its written submission, the **Alliance for German Science Organisations** states that "preserving research data over the long term and making them available […] does not only serve the verification of prior results, but also, to a large extent, the obtaining of future ones". They add that "infrastructures are to be developed according to these requirements and, if possible, interoperability integrated in international and interdisciplinary networks from the start". **NWO** refers to the importance of further development of infrastructure. They add that the role of institutional repositories and repositories infrastructures to manage these data needs to be recognised, also referring to university libraries now moving fast to meet the needs of researchers. The **Finnish Ministry of Education and Culture** also contributed on the issue, noting that "there needs to exist a clear research data management framework, which takes into account factors such as sustainability, institutional policies and procedures, IT infrastructure […], and information infrastructure […]. **DIST-CNRS** states that research data "is all the easier to manage, share and re-use if stored in international, cooperative and sustainable research data infrastructures". The **Open Knowledge Foundation** believes there is no "one size fits all solution". It also argues that the research data infrastructure should be based on open source software and interoperable based on open standards. Furthermore, it suggests that data and metadata should be deposited in machine-readable and open formats, in a similar way to the position of the United States position.

From the **e-Infrastructures and information systems** perspective, the importance of data management is fundamental. Related topics such as bibliometric citations and digital object identifiers (DOIs) were also discussed in great detail. The **Institute for the Study of Labor** addressed the issue of data storage, questioning whether the form of storage should be central or distributed. They also said that libraries are somewhat behind as some journals require data to be deposited also. **OpenAIRE** put forward the view that data should be deposited in one place because one point of deposit allows work to be distributed. They also raised DMPs, stating that requirements on data preservation should be discussed at the project submission stage and become more detailed before the award of grant. **JISC and Knowledge Exchange** state in their submission that each dataset should be identified by a persistent identifier which ensures they will be traceable in the future. They also state that as deposit is viewed as a burden for researchers, the process should be made as easy as possible, also referring to the role data librarians and data scientists can play in supporting the researcher. **DANS** echoed the need to use persistent identifiers and also pointed out that data should be stored in "trustworthy archives".

DataCite referred to the technical aspects of data storage and referred to the Datacite network of 200 data stewards and 1.7 million DOIs. The **Alliance for Permanent Access** stated that most data is unfamiliar to most people and added that the value of most data is discovered after it is produced. They added that an objective is to make data useable by as many people as possible for as long as needed. They also added that basic infrastructure is being put in place. The view from **@mire** was that research institutions are in a unique position and that data should be stored in the institutions. The floor's attention was also drawn to the issue of forged research data. The viewpoint from **FIZ Karlsruhe** emphasised the need for trusted professional and interoperable infrastructures and services to support management of and access to research data in a sustainable way. **pro-iBiosphere** believe that the majority of data should be stored and made accessible in large repositories but that appendixes to publications are suitable for some small and derived data.

<u>Publishers</u>. eLife took the view that research data "should be made available in the place which best supports its use in a sustainable and reliable fashion", while also emphasizing the role of funders. **eLife** also stated that it is not ideal for data to be stored as supplementary data to a research paper that is published on the publisher's website, before elaborating to recommend moving towards the housing of data in dedicated repositories which they state should be "ideally specialized for specific data types and domains". **eLife** also underlined the importance of "effective and persistent citation systems". **STM** state that they believe it is important that authors of research publications "include identifiers and links to available underlying source data and that reversely, data centers include links from the data to all relevant publications in which these data are mentioned". They added that they are encouraging their members to make such links visible and easily accessible. They further state that reliable and trustworthy data centers are an indispensable entity in this information chain. **Elsevier** also addressed the issue of data storage in their submission, referring to the need for data cataloguing where data might not be readily available. It is argued that, in this way, users know that the data is there, but the details of the data can be protected.

<u>Libraries</u>. LIBER took the view that research libraries should reinvent themselves, adding that libraries can also curate research data and output. There was attention drawn to the skills gap amongst academic support services. The joint position paper submitted by **LIBER, OpenAIRE and COAR** states that all research data has to be registered and deposited into at least one Open Data repository. They also support the provision of data to peer reviewers as part of the peer review process. The **European University Institute** said that libraries have an important data management role to play, due to experience with collection policy. The **Max Planck Digital Library** added that the continuous monitoring of datasets would decide if the set should be kept, noting that publishing data means additional work. A position paper submitted by **five Dutch universities** (TU Delft, Erasmus Universiteit Rotterdam, TU Eindhoven, Universiteit Twente, Universiteit Leiden) outlines support for common standards and provisions which should be set for research data management and the re-use of data. The **German National Library of Medicine** referred to the contrast between natively digital material and printed material and how this can be digitised going forward. In relation to accessibility, digital object identifiers were also raised. The **University of Verona**, in its submission, described its open archive and referred to the institutional repository being built through the harvesting of the U-Gov Research Repository.

Other Voices. ICORE made the point that "there should be centralized & de-centralized storage options but more important is the usage of standardized metadata". In its submission, **ICORE** made particular reference to the new international ISO metadata standard ISO/IEC 19788. The **EPHA** hopes for a Connecting Europe Facility that can help with data management infrastructures. In their submission, **Creative Commons** took the view that "data should be stored in a publicly accessible data repository such as Dryad, Zenodo, or figshare**"**. They also suggested that "data could also be stored in institutional or disciplinary repositories, as well as project-specific repositories". They stressed the importance of these repositories being "discoverable, accessible, and have clear terms of use, both generally for the repository itself as well as for each of the data sets that may be downloaded from the repository".

5) How can we enhance "data awareness" and a culture of sharing?

Data awareness and a culture of sharing were recurring themes of discussion throughout the day. They were often interwoven with the discussion relating to previous questions, such as data accessibility and restrictions were also applicable to this debate. It is clear that this question is viewed by stakeholders as one of the most important (if not *the* most important) to consider in formulating any policy on open access to research data. At the same time, many participants stated it is also a very complex and challenging area to address.

Researchers. TNO put forward the suggestion that specific budgets should exist to assist people in sharing data. The **Istituto Superiore di Sanità** stated that all stakeholders should be involved in addressing this issue, and reiterated the importance of sharing. It was also mentioned that data should be shared in an aggregated form. **CERN** spoke of the culture of sharing as a central aspect. The key principle, as found by the ODE project, is to build social system for sharing. It is important for anyone who has data to share to be able to do so and be recognised for it. The **Academic Medical Center / University of Amsterdam** stated that data management and data sharing plans should be required. This was reiterated by the **Helmholtz Association**, which suggested that data management and sharing plans be included in applications for funding, also adding that it would be very helpful if the Commission could coordinate such processes. The **Helmholtz** Association also took the view that incentives are crucial to building a culture of data sharing, and that data should be assessed as a product of research just like journal articles. **EMBL-EBI/ELIXIR** also raised the point of requiring research consortia to include data management and sharing plans in the application process. **LIBER** stated that researchers need to be convinced that their data is worth sharing.

Industry. There was some criticism of the idea that only a culture of data-sharing needs to be spread. Industry stated that it may be just as important to promote a culture of protecting ideas that can then be transformed into commercial products. **Philips** stated that a very clear distinction needs to be made between open access to publications and open access to data. While there is broad support for publications, the view is that data should be treated on a case-by-case basis and that the decision on whether it should be shared relies on the project participants. The **Federation of German Security and Defence Industries** argued that general open access to all sorts of data cannot be in the interest of industry and that release of industry owned data to third parties should require industry's agreement. In contrast, industry also has an interest in open access to purely publicly funded research data. During the ensuing discussion, the question of how to deal with public-private partnerships and the data they create was raised by **DANS**. **LIBER** once again mentioned that the default position should be that data is open and stressed the difference between pure research and applied research.

Research Funders. Research Councils UK mentioned the principles devised by RCUK on the management of data, and reiterated that research data from public funding are a public good. **RCUK** also stated that it was very important that data should be discoverable, although there might be legal, ethical or commercial constraints on the release of data in order to ensure that the research process is not damaged by inappropriate use of the data. The **Wellcome Trust** set forth that as public research funders, their preference is that publications and data are available, and generate real opportunities. They acknowledge that funders have an important role to play in implementing policies and guiding a sustainable culture of data sharing. The **Swedish Research Council** stated that research done through universities is considered public information and there is already a Swedish bill for making public information open. The **Alliance of German Science Organisations** made the point that an appropriate range of training and support services for professional data management must be made available. **Health Action International Europe** raised the point of needing to think about incentives for companies that also share their information.

In the written submissions received, the data awareness and a culture of data are viewed as extremely important. **NWO** state "forging partnerships between funders, research community and other key stakeholders is essential for open research data policies being created on national levels and in Europe". The **Finnish Ministry of Education and Culture** state that we "need a collective will" in order to make sure "that data resources are widely available to the use of the entire society". The **National Research Council of Italy** mentions that to share "does not always mean to give for free". **DIST-CNRS** referred to the reinforcement of rewarding measures for researchers as well as developing initial and continued training for supporting staff in charge of data curation and exploitation. The **Open Knowledge Foundation** states that "academics, research institutions, funders and learned societies all have significant responsibilities in developing a culture of data sharing". It also states that funding organisations dispersing public funds have a central role to play. Also, it underlines that there is a widespread perception among scientists that sharing data may be detrimental to career development because of the current incentive system. As such, it believes that educational and promotional activities should be set up to promote open access and "disentangle myths" and encourage them to "self-identify as supporting open access".

Information Systems & e-Infrastructures. The **Institute for the Study of Labor**, which compared sharing to donating, said that sharing is a good thing but that not all researchers will share. A further point was that saving and sharing data requires a lot of work and is thus costly, and that opening data does necessarily mean sharing data. **OpenAIRE** set forth that a culture of sharing should be user-focused. **DANS** specifically referred to the DMP in the context of this question. **JISC / Knowledge Exchange** pointed out that this is really about the policy behind DMPs but also raised the issue of costs associated with implementing a DMP. In their written submission, they also point out that researchers at present do not have sufficient incentives to share their data.

The **Alliance for Permanent Access** said that we need to approach data in a way that permits us to preserve and enhance value in the most general way. **DANS** suggested that DMPs should be required in research proposals, and that particular attention should be paid to data access during and after a project. An additional point was made that data management should be eligible for funding. There was also the suggestion of acknowledging data by giving credit to data sharing by promoting data citation. **pro-iBiosphere** makes a number of suggestions on enhancing data awareness and a culture of sharing. They suggest the provision of tools that make data sharing and re-use of data easy. They also suggest adapting copyright legislation to the needs of research and the use of common standards for resource identification. They also refer to the need for journals to find methods of citation that are helpful to curators and acknowledge authors and institutions. They also favour the interlinking of datasets, metadata and publications.

Publishers. **eLife**'s view was that publicly funded research data is a public good and should be shared effectively to maximise the benefits that arise from the public funding of research. Data should be accessible, legally usable and technically usable to maximise the benefits of sharing. **eLife** supports the use of CC0 and CC-BY licences and added that funders need to act explicitly to demonstrate that they value data sharing, and explained that this could be achieved by acting as exemplars of best practice in sharing their own data, supporting those that demonstrate and embody best practice in data sharing, developing tools that support data sharing, and making data sharing a condition of funding. He added that funders should lead by example with respect to data sharing by sharing their own data effectively and efficiently. **Wiley** addressed the issue of what makes a repository trustworthy and building trust through a version of peer-review for data. They also referred to the discoverability of data. **STM** said that, for the purpose of sharing, data should be integrated with publications, peer review and bibliometrics. Their written submission also states that "it is no secret that among researchers there is at times serious concern for sharing their data, certainly if done too early before proper processing and validation of the data has taken place". They also state their keenness "to contribute and to play a supportive role in the aim of achieving better data sharing practice". **Elsevier** said that data sharing is not the same as data storing, annotation and curation. It also advocated the creation of data catalogues so there is awareness that it exists even if it is not open. There was also a suggestion of a cross-disciplinary and multi-stakeholder network for sharing best practices to exchange success stories in promoting data sharing.

Libraries. **LIBER** said that libraries should reinvent themselves and that libraries can also support the drafting of DMPs and help describe data. It also described the role libraries can play in the curation of research data and output. In their joint statement, **LIBER, OpenAIRE and COAR** refer to a data sharing ecosystem that "must be stimulated and nurtured over the coming years". **Max Planck Digital Library** said that publishing data often is perceived as additional work and unwanted discussions over data quality, and that this perception needs to change. The **TU Delft Library** said that for universities, the utmost importance is that science is given back to society.

Other Voices. Creative Commons addressed the culture of sharing in their submission. This stated that they encourage funding agencies to require that all research products resulting from projects funded with their support be made available openly and freely. This includes providing funding for making research data available openly. They also suggest that universities and research institutions consider data products which are made openly and freely available when making decisions on promotions and recognition. They also made reference to awareness and incentives. This included encouraging the "use of alternative metrics indicating awareness, dissemination and use of research products" as well as showcasing research made possible by open data. **ICORE** made the point that a culture of sharing can be developed "through development, approval and publication of European policies and the collection of good practices as well as through templates for legal issues". The main legal issues to tackle are copyright and licensing. He added that this can be further enhanced through the support of European and international initiatives for promoting open research and its broad application.

Thanks from the Commission

The European Commission would like to express its thanks to all those who contributed to this public consultation. This report was drafted by Ivan Farmer, a trainee of the European Commission (DG RTD) and reviewed by staff in DG RTD and DG CNECT. After submission to the consultation participants, it will be made publicly available on relevant websites of the European Commission.