IJDC | *Conference Paper*

FAIR Forever? Accountabilities and Responsibilities in the Preservation of Research Data

Amy Currie
Digital Preservation Coalition

William Kilbride
Digital Preservation Coalition

Abstract

Digital preservation is a fast-moving and growing community of practice of ubiquitous relevance, but in which capability is unevenly distributed. Within the open science and research data communities, digital preservation has a close alignment to the FAIR principles and is delivered through a complex specialist infrastructure comprising technology, staff and policy. However, capacity erodes quickly, establishing a need for ongoing examination and review to ensure that skills, technology, and policy remain fit for changing purpose. To address this challenge, the Digital Preservation Coalition (DPC) conducted the FAIR Forever study, commissioned by the European Open Science Cloud (EOSC) Sustainability Working Group and funded by the EOSC Secretariat Project in 2020, to assess the current strengths, weaknesses, opportunities and threats to the preservation of research data across EOSC, and the feasibility of establishing shared approaches, workflows and services that would benefit EOSC stakeholders.

This paper draws from the FAIR Forever study to document and explore its key findings on the identified strengths, weaknesses, opportunities, and threats to the preservation of FAIR data in EOSC, and to the preservation of research data more broadly. It begins with background of the study and an overview of the methodology employed, which involved a desk-based assessment of the emerging EOSC vision, interviews with representatives of EOSC stakeholders, and focus groups with digital preservation specialists and data managers in research organizations. It summarizes key findings on the need for clarity on digital preservation in the EOSC vision and for elucidation of roles, responsibilities, and accountabilities to mitigate risks of data loss, reputation, and sustainability. It then outlines the recommendations provided in the final report presented to the EOSC Sustainability Working Group.

To better ensure that research data can be FAIRer for longer, the recommendations of the study are presented with discussion on how they can be extended and applied to various research data stakeholders in and outside of EOSC, and suggest ways to bring together research data curation, management, and preservation communities to better ensure FAIRness now and in the long term.

Received 06 April 2021 – *Accepted* 20 April 2021

Correspondence should be addressed to William Kilbride, Digital Preservation Coalition, 11 University Gardens, Glasgow G128QH. Email: william.kilbride@dpconline.org

An earlier version of this paper was presented at the 16th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Digital preservation is a fast-moving and growing community of practice of ubiquitous relevance, but in which capability is unevenly distributed.[1] Capacity erodes quickly, establishing a need for ongoing reconnaissance to ensure skills, technology, and policy remain fit for changing purpose. In open science and research data communities, digital preservation closely aligns with the FAIR Guiding Principles of findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016).

There have been substantial and early contributions to the field of digital preservation from the research data community, not least through investments of the European Commission. The European Open Science Cloud (EOSC) initiative, which has recently embarked on a new phase of development, has extensively worked to promote and enable access to open science data with the stated aim of ensuring that researchers can maximize the value of their research processes and shared large-scale Research Infrastructures (RIs).

EOSC has a compelling and ambitious prospectus for the EOSC system as a Web of FAIR Data and Related Services for science: to bring scientists and their audiences together; to federate existing infrastructures; to augment these infrastructures with new added-value services; and to revolutionize research and how scientific knowledge is created in all disciplines, in all geographies (EOSC, 2020b). The EOSC is an emergent entity currently in development, with the preparation of the three layers for the first iteration of the Minimum Viable EOSC (MVE) underway: a federating core (EOSC-Core); the federation of existing and planned research data infrastructures; and a service layer comprising common services and thematic services (EOSC-Exchange).[2]

Within EOSC's development, there been advancements towards managing and storing large quantities of open data with high-performance storage.[3] However, there remains a need for ongoing investigation of digital preservation capacity as researchers, practitioners, and experts from many sectors and industries across the digital economy expand and build capabilities, not always with a common purpose or shared vision for the future. This is the dynamic flow into which EOSC's Web of FAIR Data and Related Services steps into; and was part of the FAIR Forever study's impetus.

From August to December 2020, the Digital Preservation Coalition (DPC)[4] conducted the FAIR Forever study, commissioned by the EOSC Sustainability Working Group[5] and funded by the EOSC Secretariat Project[6], to assess the current strengths, weaknesses, opportunities and threats to the preservation of research data across EOSC,

[1] In this paper, 'data preservation' and 'digital preservation' are used synonymously. We define digital preservation broadly to include all the managed activities necessary to ensure access to digital materials for as long as necessary, including changes in technology, policy and user requirements.

[2] About the European Open Science Cloud (EOSC): https://eosc-portal.eu/about/eosc

[3] Of particular note are the services being developed through the Archiving and Preservation for Research Environments (ARCHIVER) project: https://www.archiver-project.eu/about

[4] The Digital Preservation Coalition (DPC), established in 2002, is a global not-for-profit membership organization that aims to ensure a secure digital legacy by enabling its members to deliver resilient long-term access to digital content and services, and helping them to derive enduring value from digital assets and raising awareness of the strategic, cultural and technological challenges they face: https://www.dpconline.org/

[5] About the EOSC Sustainability Working Group: https://www.eoscsecretariat.eu/working-groups/sustainability-working-group

[6] The EOSCsecretariat.eu list of approved co-creation activities: https://www.eoscsecretariat.eu/funding-opportunities/list-approved-co-creation-activities

and the feasibility of establishing shared approaches, workflows and services that would benefit EOSC stakeholders. This paper draws from the FAIR Forever study's findings to document the main strengths, weaknesses, opportunities, and threats to the preservation of FAIR research data at the outset of the EOSC Association, and to explore how they can apply to preservation in the research community more broadly.[7]

The paper begins with a background of the study and overview of the methodology involving three stages of research: a desk-based assessment of the emerging EOSC vision; interviews with representatives of the European Strategy Forum on Research Infrastructures (ESFRI) Cluster and Regional projects as key EOSC stakeholder groups; and focus groups with digital preservation specialists and data managers in research organizations. The authors will then summarize four key findings from the study arranged by themes: preservation and the FAIR principles; implicit meanings and assumptions about digital preservation in the EOSC vision; need for elucidated preservation roles, responsibilities, and accountabilities; and risks to data loss, reputation, and sustainability (Currie and Kilbride, 2021). To better ensure that research data can be FAIRer for longer, the final section outlines the recommendations presented to the EOSC Sustainability Working Group with discussion on how they can be extended and applied to various research data stakeholders in and outside of EOSC, and suggest ways to bring together research data curation, management, and preservation communities to better ensure FAIRness now and in the long term.

Background and Methods

The EOSC Sustainability Working Group was established to develop and provide a set of recommendations concerning the implementation of an operational, scalable and sustainable EOSC federation after 2020, one that will gradually open up its user base to the public sector and industry. This resulted in the Solutions for a Sustainable EOSC report (EOSC Sustainability Working Group, 2020), taking into account feedback provided by over 30 EOSC projects, organizations, governance and executive board members, other working groups, and integrating the input from several studies commissioned by the Sustainability Working Group. [8]

The long-term preservation of FAIR research data was identified as a critical area of sustainability. For this reason, the DPC was approached by the Sustainability Working Group in June 2020 to submit an EOSCSecretariat.eu co-creation study proposal to conduct a study on long term data preservation roles and responsibilities for EOSC. The objective of the study was to provide information and recommendations to the Sustainability Working Group by assessing the current strengths, weaknesses, opportunities and threats to the preservation of research data across EOSC, and the feasibility of establishing shared approaches, workflows and services that would benefit EOSC and its immediate stakeholders.

The DPC was well suited for the aims of the study. It is a global not-for-profit membership organization dedicated to addressing digital preservation. It has a large network of like-minded organizations supported by established social and organizational infrastructure. With existing and effective communication channels in place, as well as a strong web and social media presence, the Coalition draws from the expertise of both its staff and diverse membership in developing and sharing approaches and workflows for

[7] EOSC Association: https://eosc.eu/

[8] EOSC Sustainability Working Group: https://www.eoscsecretariat.eu/working-groups/sustainability-working-group

preservation. DPC is also vendor and technology-neutral, offering clear and impartial advice on the maturity and suitability of different digital preservation tools, products and techniques.

The FAIR Forever co-creation study proposal was submitted to the EOSC Secretariat in June 2020 and approved by the EOSC Secretariat in August 2020, with the study commencing on 18th August 2020.

Research Design and Methods

The initial proposal was for the DPC to undertake a study on the current strengths, weaknesses, opportunities and threats to the preservation of research data across EOSC, and the feasibility of establishing shared preservation approaches, workflows and services for EOSC stakeholders. The study used a team-based approach and a lightweight agile methodology. The authors of this paper were the two primary researchers, but the study drew on the experience and expertise of other DPC team staff members contributing to main project tasks as appropriate. Frequent project meetings and check-ins with the Sustainability Working Group were established to allow effective coordination and application of team skills to the activities at hand. This was designed to ensure the most effective use was made of the available project resources to deliver high quality and insightful project results. Where appropriate, the project team offered to involve DPC members and various EOSC stakeholders in reviewing documents, ensuring their participation from the outset and enhancing the likelihood of their subsequent buy-in.

Three distinct phases of research were established, initiated, developed, and executed for the study, with each outlined in the following subsections and detailed further in the final report (Currie and Kilbride, 2021).

Desk-based assessment

The study began with a desk-based assessment to establish ‘state of the art’ in digital preservation thinking within the EOSC stakeholder community. In other words, to review EOSC governance documentation, EOSC projects’ outputs and plans, and other relevant literature addressing the current state of development of digital preservation approaches, workflows, and services for the envisioned EOSC Web of FAIR Data and Related Services as communicated in those materials.

In this way, the desk-based assessment involved a short but intensive phase of qualitative research and data collection, undertaking a review of 20 documents from 18th August to 1st September 2020. This initial assessment involved multiple readings of the provided documents to identify, collect, and assess explicit or implicit references to areas of technological, organizational, and policy directly or indirectly relating to digital preservation roles, responsibilities, and activities.

The reviewed documents for the two-week desk-based assessment were those provided by the Sustainability Working Group. However, over the course of the study, the researchers also collected and reviewed information from updated versions of EOSC governance documents, resources provided or suggested by participants or the Sustainability Working Group, and from materials, events and activities such as the EOSC Governance Symposium (Currie and Kilbride, 2021). The constant throughput of documents, especially the Strategic Research and Innovation Agenda (SRIA), meant this was a continuing task throughout the research. [9] Upon project completion in December 2020, there were over 40 items reviewed over the course of the study.

There was no uniform set of codes applied for analysis of each of the documents, but rather keywords (data preservation, digital preservation, archiving, long-term) and emerging themes (data management planning, community or bottom-up archiving, storage and computing, data types, FAIR principles, skills and training, funding) from the readings. These themes were used for subsequent reading and analysis by the two primary researchers, with those findings reviewed by other DPC staff members for refining and sense checking, with the refined findings presented to the Sustainability Working Group for further feedback or comment.

Part of the analysis included the identification of how the EOSC stakeholder community was defined within the documents, and the analysis found consistency among the documents in the grouping of stakeholders into the following groups by the EOSC Governance[10]

- EOSC Governance Board: representatives from EU countries, countries associated with Horizon 2020 and the Commission to ensure effective supervision of the EOSC implementation.
- EOSC Executive Board: representatives from the research and e-infrastructures communities.
- EOSC Stakeholders: a wider range of actors, consulted through a series of stakeholder events and online consultations to collect input and recommendations.

These general groups helped structure the three stages of research outlined in the proposal. Whereas the desk-based assessment focused on collecting information communicated in higher-level EOSC Governance documents, the other two stages of research each focussed attention on collecting information about preservation communicated by representatives from the other two groups.

Semi-structured interviews

The second stage of research collected data from semi-structured interviews with representatives of the researcher and e-infrastructure communities, specifically the European Strategy Forum on Research Infrastructures (ESFRI) Cluster and Regional projects. The interviews were used to test and refine findings from the desk-based assessment, recognizing these findings were drawn from qualitative analysis of limited information surrounding digital preservation provided in the reviewed documents and to allow other findings to emerge through discussions of their current requirements and capabilities in digital preservation.

DPC chose a semi-structured approach to elicit contributions and reflections from participants as EOSC stakeholders. The interviews were structured around a consistent set of themes drawing on initial findings from the desk-based assessment, but with mostly open-ended questions that encouraged flexibility and allowed other relevant

[9] The study reviewed multiple versions of the EOSC SRIA including the Open Consultation version (20 July 2020), Version 0.8 (18 October 2020), and Version 0.9 (16 November 2020). The most recent version 1.0 (15 February 2021) is available at the EOSC Association website: https://eosc.eu/

[10] Interim EOSC Governance 2018-2020: https://www.eoscsecretariat.eu/eosc-governance. In July 2020, the EOSC Association was set up to provide a single voice for advocacy and represent the broader EOSC stakeholder community, see: https://eosc.eu/join-association

ideas around the topics to emerge during discussions. Interview guides were sent to participants in advance of the interviews to provide an overview of the study, the scope and structure of the interview, and an outline of the main topic areas and questions to be covered in case they wanted to give written answers or verify factual matter.[11] In this way, the interview guides provided an interview instrument for gathering data with uniform topic areas offering a set of codes for collecting and analyzing data.

The interviews were one hour in length, conducted in English, and used the same web-based video platform, with two DPC staff interviewers and either one or two interviewees attending. Invited participants were encouraged to review the questions provided in the interview guide and to suggest or bring in other representatives from their ESFRI stakeholder groups to respond to the different topic areas. One of the researchers took the role of interviewer and the other as designated notetaker. Interviews were recorded with permission and reviewed to confirm and supplement the notes taken.

In total, seven interviews were conducted with twelve individuals from the following ESFRI stakeholder groups: NI4OS-Europe, ESCAPE, SSHOC, PaNOSC, ENVRI-FAIR, and EOSC-Life.[12] Representatives from EOSC-Pillar, EOSC-Nordic, and EOSC-Synergy were contacted but were either unavailable or unresponsive to the emailed requests during the interview period of the study.[13] These interviews occurred from the 6th October to 2nd November 2020.

Focus groups

The third stage of research engaged with representatives of the research data management and digital preservation community to help articulate, assess, and compare potential use cases for preservation services within EOSC. The researchers purposively selected and invited participants meeting the following criteria to gather specialized expert knowledge and practical experiences with – and to some degree biases toward – digital preservation: individuals at DPC Member organizations, who work at a Research Performing Organization (RPO), are based in Europe (DPC's membership is predominantly United Kingdom and Ireland), and their role is in the area of research data management or digital preservation.

The rationale for the narrowed selection criteria for participants followed critical case thinking in the sense that if challenges occur for this group, then it is very likely that other, less knowledgeable individuals at less developed organizations will face the same. Furthermore, it explicitly highlighted those in the digital preservation community as part of the wider range of EOSC Stakeholders. The focus group activities provided a means for consultations to collect input and recommendations from those in the digital preservation community to challenge assumptions implicit within EOSC more broadly.

A broad definition of a use case was adopted for discussing, constructing, and analysing possible use cases scenarios within the focus group session. A use case was broadly understood as a situation in which a resource or service could be used to support the daily work of those managing and preserving research data at RPOs. The decisions to adopt a broad definition of a use case and purposive sampling and were developed in light of the emerging findings from the interviews, ongoing developments in archiving and preservation services within EOSC, and revised interim statement. The complexity

[11] A copy of the generic interview guide used as the interview instrument, with questions arranged by topic area, is available in the appendix of the final report (Currie and Kilbride, 2021).

[12] A list of the participants, with names anonymized, is available in the appendix of the final report (Currie and Kilbride, 2021).

[13] The authors welcome and encourage any feedback or contribution from these groups to this article or the final report.

in different user needs and requirements among the range of EOSC stakeholders is well established from EOSC research projects, and touched upon through the discussion with the stakeholders in interviews. Rather than add to or reiterate this complexity, the study narrowed its scope for the identification of shared goals or challenges among a key group of stakeholders working in research data management and preservation activities at RPOs.

A focus group session was one hour in length, conducted in English, and used the same web-based video platform. The researchers wanted to encourage participants to speak freely, so we chose not to record the workshop sessions but instead collected data through designated notetakers, group activity sheets, and responses to an online survey. Each session followed the same general structure: a brief overview and background of EOSC, FAIR, and the study; a group activity (or walk through) discussing potential use case scenarios based on representative personas (actors) within the digital preservation community; reporting key points from activity discussion, and more general discussion touching upon the topics or areas surrounding findings and interim statement; and survey questions incorporated throughout to supplement and facilitate discussion.

In total, there were fifteen participants and four focus group sessions; a group session with ten participants and three follow up interactive sessions with five participants who were unable to attend the group session. These interactions ran from the 9th November to 2nd December 2020.[14] Analysis included the qualitative and quantitative data collected during the sessions. However, the quantitative data collected through the survey was used to supplement the qualitative data; it was intended to support logical deductions from the analysis rather than provide generalizable findings.

Note on Study Execution and Constraints

In addition to the delimitations noted in the research methods, there were acknowledged limitations of the study. Although the activities and deliverables within the study largely followed the initial specification, the timeline of the study was significant impacted by a late start. Initially proposed in June, the study was not commissioned till mid-August. A simple rectification was to move the key milestones out by three months but, to synchronize with key milestones on the roadmap for the EOSC Strategic Research and Innovation Agenda (SRIA), there was the need to compress deadlines leaving little time to digest and elucidate findings.

As a result, preliminary findings were presented iteratively to meet the timelines of the Sustainability Working Group, who in turn were responding to a demanding but rigid programme. Follow up interviews and questions were curtailed, and peer review of emerging findings has been scaled back to ensure that deadlines were respected. A benefit from this iterative approach was that it allowed for continuous communication with members of the Sustainability Working Group, challenging, refining and informing the research throughout. The project team gratefully acknowledges the generous support of the Sustainability Working Group to this research and also acknowledges the timeline as a constraint on the research.

[14] An overview of the survey findings about the focus group participants, as well as a list of participants with names anonymized, is available in the final report (Currie and Kilbride, 2021).

Key Findings

The research methods employed in the study were largely qualitative by design to meet the overall aims and purpose of the study; to provide information and recommendations to the Sustainability Working Group by assessing the current strengths, weaknesses, opportunities and threats to the preservation of research data across EOSC, and the feasibility of establishing shared approaches, workflows and services that would benefit EOSC and its immediate stakeholders. This is why the key findings from the study, presented in the next sections, are not intended to be generalizable but rather transferable; the findings emerging from the different qualitative methods were presented iteratively to members of the Sustainability Working Group, challenging, refining and guiding their transferability in the context of sustainability and FAIR data preservation in the context of EOSC (Currie and Kilbride, 2021).

FAIR for Now? Preservation and the FAIR Principles

From the outset of this study, it has been apparent that EOSC and the FAIR principles are tightly interdependent. The FAIR principles were part of the foundations of the envisioned EOSC, enabling researchers to perform Open Science and open their research data for sharing (EOSC, 2020b). The successful federation of research data infrastructures for EOSC requires the implementation of the FAIR guiding principles so that the data and digital content is discoverable and usable. The EOSC FAIR Working Group, FAIR task groups, and other related initiatives and projects have worked to define and communicate the corresponding FAIR requirements and practices expected for EOSC stakeholders.[15]

There are significant areas where the FAIR principles intersect with preservation, and there are notable examples of how good practice applying FAIR principles also delivers good practice in digital preservation. These include an early focus on persistent identifiers (PIDs), an emphasis on data management planning, and planning for robust storage. As the study progressed, further plans for repository audit and certification became apparent (EOSC FAIR Working Group, 2021).[16] All of these are essential elements of a digital preservation strategy and capability.

There are also areas where the implementation of FAIR principles falls short of what might be achieved, even when they align with an existing culture and expectation. For example, while there is a general recognition of the value of researchers creating and submitting data management plans (DMPs), and even a sense that responsibilities are made clear through the process, interviewees doubted whether they delivered the impact which was intended. Also, the strengths and limitations of FAIR for creating, maintaining, and preserving metadata also came up in conversations. When asked about FAIR during the interviews, one participant mentioned insufficient ways to automatically preserve metadata for findability. Another interviewee ranked interoperability as the greatest of the FAIR challenges – it was far more difficult to make data interoperable than findable or accessible. When asked about FAIR interoperability for archiving and preserving, he echoed the constraints of time and resources. An interviewee working with humanities data also commented that interoperability over time depends on data being FAIR at the outset and that the success will come from re-

[15] EOSC FAIR Working Group: https://www.eoscsecretariat.eu/working-groups/fair-working-group

[16] See also FAIRsFAIR Support programme for Data Repositories: https://www.fairsfair.eu/application-results-open-call-data-repositories

use as the 'proof of the pudding'. A degree of hesitation expressed by the participant of wanting to avoid 'FAIR fatigue' – worrying that these efforts without concrete embodiment of the EOSC vision of FAIR data will fall through.

A similar pattern of aspirational versus achievable goals for preservation emerged with respect to the audit and certification of repositories. This is a welcome step and is articulated within the EOSC vision (EOSC, 2020c): but, as currently envisioned, it is insufficient to the scale of the challenge if taken in isolation because it gives only partial consideration of the path dependency associated with digital preservation actions required. One participant anticipated that the EOSC interpretation of FAIR meant that repositories would likely be presented with data that they had limited resource to preserve, and little practical chance of saving.

In summary, while the findings showed the benefits of EOSC developing FAIR among services and areas where the implementation of the principles might benefit both data management and preservation planning, there seems to be a gap in EOSC's interpretation of the FAIR principles as they pertain to preservation. As currently articulated, EOSC's implementation of FAIR – notably through the work and contributions of FAIRsFAIR – helps researchers and data managers assess their awareness of the requirements for making data FAIR prior to uploading them into a repository, but it does not provide a consistent programme for preserving data.[17] The implementation of FAIR principles in general and specifically concerning digital preservation appears to be ultimately in the hands of those providing and managing the data infrastructures, with some of them commenting on limited time and the current challenges of interoperability.

Digital Preservation is not Explicit: The Meanings of Digital Preservation in the EOSC Vision

Whilst there were references to the preservation and archiving of research data throughout the governance documents reviewed for the study's desk-based assessment – notably in the Strategic Research and Innovation Agenda (SRIA), published on 20th July 2020 – closer readings and analysis of these documents found that digital preservation is only implicit in the EOSC vision (EOSC, 2020a). There were implicit meanings and assumptions about digital preservation – and data – in the EOSC vision and among stakeholders. Hence, there was a need to test and refine this initial finding from the desk-based assessment through the interviews to determine whether preservation is also implicitly understood by them as identified EOSC stakeholders.

In general, most agreed that there is an overall need for a more explicit digital preservation policy and strategy. There is an acknowledgement within the EOSC vision that preservation is important and a consensus that it is a core requirement in every discipline: but specific preservation functions remain obscured and miscommunicated. When asked how digital preservation might be made more explicit in the EOSC vision, one interviewee suggested that an articulation of the main digital preservation objectives and challenges could help guide and assess research infrastructure requirements and capabilities. Another participant added that she felt it was critical for them (as representatives of stakeholders groups and researchers) to be involved with or aware of how EOSC will establish requirements for preservation within the existing policies and frameworks.

In summary, a clearer articulation of data in digital preservation within EOSC's strategic mission, along with an effort to spell out objectives, challenges, and implications

[17] FAIRsFAIR: https://www.fairsfair.eu/

for the preservation of research data will help strategic alignment in and across EOSC infrastructures. Stakeholders should be aware of and recognize the width of the preservation challenge implied by a broad, maximal definition of data; data sets, publications, correspondence, software, applications, libraries, code, micro-service dependencies, execution environments and operating systems, which will all need to be preserved or recreated depending on scientific use cases.[18]

The Need for Elucidation: Preservation Roles, Responsibilities, and Accountabilities

The most significant findings from the desk-based assessment and interviews concerned preservation roles and responsibilities that are unclear and accountabilities that are uncertain. These findings lead to a significant number of questions about the configuration of digital preservation capability not just in EOSC but in the larger research data management and research communities.

There was a perception by some stakeholders that EOSC would provide or fund end-to-end digital preservation solutions. For example, the following stated in the September 2020 'ExPaNDS and PaNOSC position paper on EOSC: A communication following the SRIA consultation':

> 'The data produced by our instruments at PaN facilities are a valuable resource for long term reuse. However, our data policies typically guarantee only a 5-10 year preservation period. After that, it would be valuable if the EOSC would take over the responsibility for finding facilities/services for storage and curation of FAIR data to keep it available for further long-term use' (ExPaNDS and PaNOSC, 2020).

When interviewees were asked about the present view taken by the EOSC Governance Board that the EOSC community is ultimately responsible for the preservation of research data, the participants largely accepted this responsibility but noted that within that broad community, there are different views of who should be responsible. The general sense was that researchers are responsible for data creation and management until the data is stored or transferred to repositories. Interviewees discussed the practical limitations of this approach for ensuring the quality of research data and compliance with data management plans. One explained that cluster projects could provide high-level guidance, but data responsibility sits with the partners and infrastructures. Another argued that "data buckets are not preservation functions, and long-term preservation needs domain knowledge" – because domain knowledge cannot be recreated afterwards.

The concept of data stewardship was often framed as an ambassadorial role between the researcher and other stakeholders. The SRIA Version 0.8 noted "When open science becomes the 'new normal', scientists will extend their requirements accordingly, and new roles and responsibilities will have to be created (e.g. data scientists, data stewards, etc.)" (EOSC, 2020b). Those interviewed, who are working in EOSC FAIR initiatives and projects, saw data stewardship as a critical way to support interoperability. For example, the FAIRsFAIR project 'Recommendations on practice to support FAIR data principles' (2020) outlines specific recommendations aimed primarily at research communities and research support personnel including data stewards and

[18] For example, software in particular is a significant digital preservation challenge as the certification of code repositories and the validation of emulation or virtualization services are still immature.

research software engineers. Yet, it is unclear if such stewardship roles are also professionally responsible for long-term preservation as the title implies. Moreover, if they are assumed to hold this responsibility for long-term preservation, questions arise to about how it will be supported in the long term.

Consequently, there is a lack of clarity on preservation responsibility or the skills needed to guide data creators. One participant noted that researchers in smaller social sciences and humanities institutes were willing to share data but are unsure how to find and navigate the best strategies for preserving it; and that researchers had a greater preference for giving their data to institutional repositories because they believe it could ensure better data preservation as there is staff with a job to curate and control the data and, should anything go awry, will be held accountable.

Conversations with the interviewees and later interactions with those in the digital preservation community highlighted that the problems arise not in the unwillingness of those in institutional repositories to take this challenge on, but rather the staffing and resource limitations that are implied. There is a need for digital preservation skills development and training for staff. The interviewees felt organizational viability across the EOSC community varied in terms of governance, organizational structure, staffing and resourcing of digital preservation activities. For example, one interviewee said that digital preservation training for their librarians at their university was needed and another said preservation skills are not well represented in the overall infrastructure, and more spend is given to equipment and instruments than people committed to taking care of the data afterwards. She added that in the more matured clusters and science communities, there are projects led by domain experts who are familiar and knowledgeable with digital preservation policy and strategies. For example, in the context of EPOCH (European Research Network of Excellence in Open Cultural Heritage), the work of national component repositories where the expertise and capacity exists.[19] Another commented that preservation is – or more accurately should be recognized as – a 'real professional' job which makes a distinct and essential contribution to the research data lifecycle. Interviewees noted that a coordinated and central approach to the provision of preservation skills should be a genuine priority for EOSC.

EOSC has a substantial community around it and relatively good infrastructures of communication. While it is agreed that collaboration is key for success, it is complicated when dealing with the scale of EOSC. Some of the issues in scoping preservation within EOSC arise because the services are ordered around community needs. This proximity to the community comes at the cost of complexity but is a very sound investment for data creators and users, creating much better preservation outcomes.

Digitally Endangered Species? Risks to Data, Reputation, and Sustainability

The research benefit to preserving data was evident to all the participants across disciplines and regions, especially when discussed in the context of making research data FAIR. Their concern was how to address other benefits of preservation and assess costs in order to secure the resources required to create and maintain a sustainable digital preservation programme.

One area where the costs of digital sustainability appeared markedly unclear was that of personnel and staffing costs.[20] Participants found the costing of preservation in

[19] The European Research Network of Excellence in Open Cultural Heritage (EPOCH) project webpage: https://www.brighton.ac.uk/csius/what-we-do/research-projects/epoch.aspx

[20] Although there are digital models and tools including the following, none were mentioned by

terms of personnel and staffing costs difficult to calculate or estimate, including one interviewee who was currently working on a project on sustainability. An interviewee commented that most of the partners in one of the ESFRI Research Infrastructures Cluster projects have no budget for preservation and questioned whether it is even possible to 'make a guess' on the costs. Others agreed that parameters and actual measures were lacking; several interviewees felt that sustainability is not guaranteed. Nevertheless, given the expressed importance of knowledge, skills, and expertise to developing and implementing a preservation programme, participants agreed that there should be deeper investigations on how to calculate the costs of preservation roles and responsibilities so that they can be included and aligned with EOSC funding programmes.

The risk of data loss or reputational harm or sustainability for EOSC arises from the implicit preservation requirements and implied responsibilities. Stakeholders recognize that data is 'born vulnerable', but there is a lack of clarity and depth of insight into how to address this issue, which creates reputational risk. This finding, presented to the Sustainability Working Group in an interim statement, was situated in the context of the DPC's Global List of Digitally Endangered Species (BitList) which notes that digital materials are 'Critically Endangered' in the presence of two conditions.

> 'Digital materials are listed *Critically Endangered* when they face material technical challenges to preservation, there are no agencies responsible for them or those agencies are unwilling or unable to meet preservation needs. This classification includes *Endangered* materials in the presence of aggravating conditions' (Digital Preservation Coalition, 2020).

This is the second-highest alert level and is a precarious position to place emerging EOSC data infrastructure.

The BitList also describes how good practice can reduce the alert level pertinent to any given set of digital materials (DPC, 2020). This is an important corollary in the context of EOSC where key services and stakeholders model good practice through policy development, training and procedural development. An opportunity exists to address these challenges if preservation requirements and accountabilities were explicit, preservation risks managed across the data lifecycle, and strategic alignment encouraged at the highest level of the EOSC vision.

Recommendations and Discussion

The findings from the three stages of data collection and analysis – from the desk-based assessment, the interviews with representatives of the ESFRI Cluster and Regional projects, and the focus group sessions on use case scenarios for preservation services in EOSC – supported that:

- Digital preservation is not explicit in the EOSC vision: it needs to be.

interviewees when asked for examples of costing models used: Collaboration to Clarify the Costs of Curation (4C) Roadmap: https://www.4cproject.eu/; Curation Costs Exchange (CCEx): https://www.curationexchange.org/; AV Preserve (AVP) Costs of Inaction Calculator: https://coi.weareavp.com/; Keeping Research Data Safe (KRDS) Benefits Framework and a Benefits Analysis Toolkit: https://beagrie.com/krds-i2s2.php

- Roles, responsibilities, and accountabilities for preservation in EOSC are opaque: they should be clarified.
- There is a risk to data, reputation and sustainability: EOSC cannot achieve its goals in the long term unless they are addressed.

From these arguments, the authors presented recommendations based on analysis of cumulative findings to members of the Sustainability Working Group in the final report, which was submitted to the EOSC Secretariat in December 2020, and published with open access in February 2021 (Currie and Kilbride, 2021).

Having identified incipient strengths and weaknesses, the final report of the study concluded with nineteen recommendations for the Sustainability Working Group to promote and consider, which were arranged by seven areas of action, and also arranged with respect to owners of the recommendation (see Figure 1).

A follow-up meeting in January 2021 with contacts of the Sustainability Working Group was conducted to discuss the reaction, feedback, and comments to the final report submitted to the EOSC Secretariat. There was an overall positive response to the recommendations with identified gaps acknowledged and selected recommendations included in the Solutions for a Sustainable EOSC (EOSC Sustainability Working Group, 2020). Furthermore, findings from the study have been included in the most recent version of the SRIA.

> 'The extent to which institutions have been given or taken explicit responsibility for preservation is unclear, assuming even that they have the capability to deliver. The concept of data stewardship at present, although it may imply preservation, is more often seen as an ambassadorial role, between the researcher and other institutional departments and staff such as the computing services, institutional repositories, libraries or archives. Clearer roles and responsibilities are needed, including the assessment of capability as well as functions, salaries and funding streams for preservation' (EOSC, 2021).

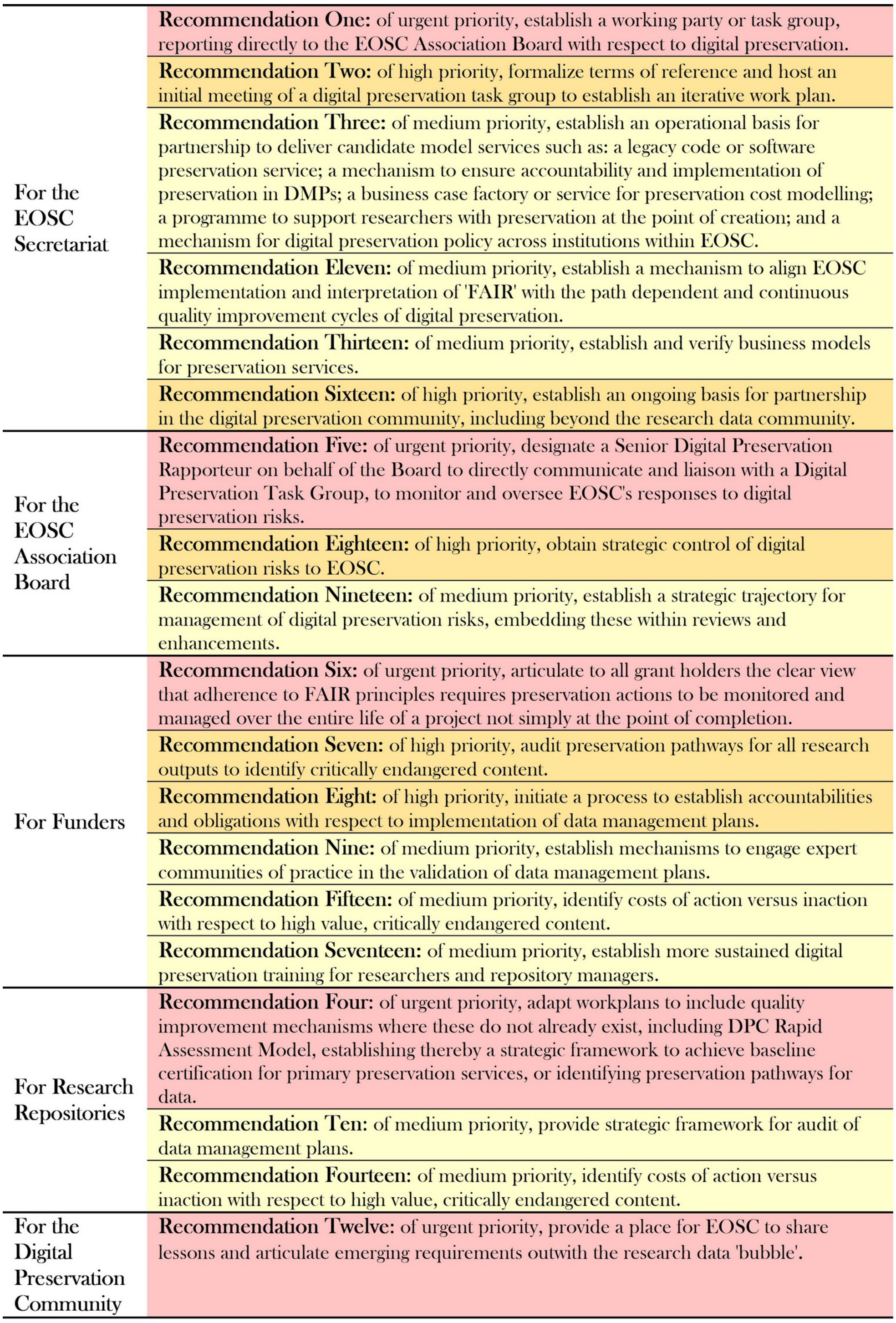
Perhaps more significant is the recent announcement that the EOSC Association is in the process of establishing a digital preservation advisory group, addressing recommendations of urgent priority (Recommendation One; Recommendation Five).[21] With the establishment of this group, there is a path for the adoption and implementation of the high priority recommendations that follow (e.g. Recommendations Two, Sixteen, and Eighteen).

Presentations of the key findings were also delivered during a 14th January Science Europe workshop on maturity matrices for research performing organizations, research funding organizations, and research infrastructures.[22]

This ongoing work of Science Europe in developing matrices for quality improvement mechanisms in research data management align closely with those for preservation (Recommendation Four), suggesting space for further collaboration and development across research data management and preservation communities.

[21] This was announced by Bob Jones, EOSC Association Director, during the March 2021 'FAIR Forever? FAIRer for longer: Digital preservation and the European Open Science Cloud' webinar hosted by the DPC: https://www.dpconline.org/events/fair-forever-event

[22] Science Europe 'Achieving Sustainable Research Data' workshop 14 January 2021.



For the EOSC Secretariat	**Recommendation One:** of urgent priority, establish a working party or task group, reporting directly to the EOSC Association Board with respect to digital preservation.
	Recommendation Two: of high priority, formalize terms of reference and host an initial meeting of a digital preservation task group to establish an iterative work plan.
	Recommendation Three: of medium priority, establish an operational basis for partnership to deliver candidate model services such as: a legacy code or software preservation service; a mechanism to ensure accountability and implementation of preservation in DMPs; a business case factory or service for preservation cost modelling; a programme to support researchers with preservation at the point of creation; and a mechanism for digital preservation policy across institutions within EOSC.
	Recommendation Eleven: of medium priority, establish a mechanism to align EOSC implementation and interpretation of 'FAIR' with the path dependent and continuous quality improvement cycles of digital preservation.
	Recommendation Thirteen: of medium priority, establish and verify business models for preservation services.
	Recommendation Sixteen: of high priority, establish an ongoing basis for partnership in the digital preservation community, including beyond the research data community.
For the EOSC Association Board	**Recommendation Five:** of urgent priority, designate a Senior Digital Preservation Rapporteur on behalf of the Board to directly communicate and liaison with a Digital Preservation Task Group, to monitor and oversee EOSC's responses to digital preservation risks.
	Recommendation Eighteen: of high priority, obtain strategic control of digital preservation risks to EOSC.
	Recommendation Nineteen: of medium priority, establish a strategic trajectory for management of digital preservation risks, embedding these within reviews and enhancements.
For Funders	**Recommendation Six:** of urgent priority, articulate to all grant holders the clear view that adherence to FAIR principles requires preservation actions to be monitored and managed over the entire life of a project not simply at the point of completion.
	Recommendation Seven: of high priority, audit preservation pathways for all research outputs to identify critically endangered content.
	Recommendation Eight: of high priority, initiate a process to establish accountabilities and obligations with respect to implementation of data management plans.
	Recommendation Nine: of medium priority, establish mechanisms to engage expert communities of practice in the validation of data management plans.
	Recommendation Fifteen: of medium priority, identify costs of action versus inaction with respect to high value, critically endangered content.
	Recommendation Seventeen: of medium priority, establish more sustained digital preservation training for researchers and repository managers.
For Research Repositories	**Recommendation Four:** of urgent priority, adapt workplans to include quality improvement mechanisms where these do not already exist, including DPC Rapid Assessment Model, establishing thereby a strategic framework to achieve baseline certification for primary preservation services, or identifying preservation pathways for data.
	Recommendation Ten: of medium priority, provide strategic framework for audit of data management plans.
	Recommendation Fourteen: of medium priority, identify costs of action versus inaction with respect to high value, critically endangered content.
For the Digital Preservation Community	**Recommendation Twelve:** of urgent priority, provide a place for EOSC to share lessons and articulate emerging requirements outwith the research data 'bubble'.

Figure 1. Nineteen recommendations tabulated by owners.

Indeed, the number of registrations for a DPC webinar on FAIR, EOSC, and the FAIR Forever study indicates there is an interest in creating opportunities – and a place – to share lessons and articulate emerging requirements outwith the research data 'bubble' (Recommendation Twelve). For this event, there were 76 registrations from 25 countries, with a mix of those from research, data management, and preservation communities.[23]

[23] For example, 35 of the registrations were DPC members, 12 EOSC Association members, 17 were

At the same time, given that the EOSC is an emergent entity currently embarking on a new phase following the completion of the first in December 2020, many of the recommendations and findings offered in the report are no longer as current. This is especially true for Recommendation Three pertaining to candidate service models, as new services and tools, particularly those being developed through the Archiving and Preservation for Research Environments (ARCHIVER) project, have made progress in recent months.[24]

There is certainly a place for deposit-storage-access systems and services that provide basic bit-level assurance, but there remains the ongoing need for data expertise and more advanced preservation activities requiring the identification and support of preservation roles, responsibilities, and activities. FAIR principles – especially interoperability – are at risk without this data expertise and support for long-term preservation actions. Some may see this conclusion as somewhat self-evident and non-controversial, but if this is so, the question arises as to why it has not been made more explicit not just in EOSC but in the open science and research data communities more broadly?

Therefore the authors wish to put forth this question to readers for further research and discussion to a broader audience of readers to solicit feedback and comment from those who identify themselves in the stakeholder groups mentioned, but also those part of the broader open science and research data communities. To better ensure FAIRness now and in the long term, are there common understandings of FAIR across these communities? For example, just as issues of interoperability for findability, access, and re-use may not be given the same level of attention as they are in the digital preservation community, there may be a need for greater attention in the digital preservation community to issues of accessibility with consideration of subsequent reusability.

Acknowledgements

The EOSC FAIR Forever study received funding from the European Union under the EOSC Secretariat project, grant agreement number 831644.

The authors gratefully acknowledge the support of the EOSC Sustainability Working Group, especially Bob Jones (CERN), Jessica Klemeier (EMBL) and Marie Timmermann (Science Europe), and the many partners and colleagues who gave freely of their time and expertise in interviews and at the focus groups.

References

Currie, A. & Kilbride, W. (2021). *FAIR Forever? Long Term Data Preservation Roles and Responsibilities, Final Report* (Version 7 17 February 2021). doi:10.5281/zenodo.4574234

Digital Preservation Coalition. (2020). *Global List of Digitally Endangered Species (BitList).* doi:10.7207/DPCBitList20-01

service provider organizations, and 27 from research performing organizations.

[24] ARCHIVER project: https://www.archiver-project.eu

European Open Science Cloud. (2020a). *Open Consultation for the Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud* (20 July 2020). Retrieved from https://www.eoscsecretariat.eu/sites/default/files/open_consultation_booklet_sria-eosc_20-july-2020.pdf

European Open Science Cloud. (2020b). *Strategic Research and Innovation Agenda (SRIA)* (Version 0.8 18 October). Retrieved from https://www.eoscsecretariat.eu/sites/default/files/eosc-sria-v08.pdf

European Open Science Cloud. (2020c). *Strategic Research and Innovation Agenda (SRIA)* (Version 0.9 16 November 2020). Retrieved from https://www.eoscsecretariat.eu/sites/default/files/eosc-sria-v09.pdf

European Open Science Cloud. (2021). *Strategic Research and Innovation Agenda (SRIA)* (Version 1.0 15 February 2021). Retrieved from https://eosc.eu/sites/default/files/EOSC-SRIA-V1.0_15Feb2021.pdf

European Open Science Cloud FAIR Working Group. (2021). *Recommendations on certifying services required to enable FAIR within EOSC* (January 2021). doi:10.277712753

European Open Science Cloud Sustainability Working Group. (2020). *Solutions for a Sustainable EOSC: A FAIR Lady (olim Iron Lady Report)* (November 2020). Retrieved from https://op.europa.eu/en/publication-detail/-/publication/581d82a4-2ed6-11eb-b27b-01aa75ed71a1

ExPaNDS & PaNOSC. (2020). *ExPaNDS and PaNOSC position paper on EOSC: A communication following the SRIA consultation* (September 2020). Retrieved from https://expands.eu/2020/09/21/expands-and-panosc-position-paper-on-eosc/

FAIRsFAIR. (2020). *D3.4 Recommendations on practice to support FAIR data principles.* doi:10.5281/zenodo.3924132

Netwerk Digitaal Erfoed (Digital Heritage Network). (2017). *Digital Sustainability Cost Model report* (January 2017). Retrieved from doi:10.5281/zenodo.4274253

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3.* doi:10.1038/sdata.2016.18