FAIR + Time: Preservation for a Designated Community

Purpose

This draft working paper has been prepared by members of the SSHOC[1], EOSC Nordic[2] and FAIRsFAIR[3] projects working on related topics. It seeks to present some key concepts and expand on them to specify the standards and assessments required for an interoperable ecosystem of FAIR (findable, accessible, interoperable and reusable) data preserved for the long term in generalist and specialist FAIR-enabling trustworthy digital repositories (TDR) for a defined designated community of users. It seeks to provide context and define these concepts for audiences familiar with research data and technical data management systems, but with less direct experience of digital preservation and trustworthy digital repositories. This is intended to help clarify which organisations are potential candidates to receive CoreTrustSeal TDR status and to identify and support the types of organisations that may not be candidates but also play a vital role in the data ecosystem.

[1] https://sshopencloud.eu/
[2] https://eosc-nordic.eu/
[3] https://www.fairsfair.eu/

Synopsis

The digital object management aspects of Trustworthy Digital Repository (TDR) standards (CoreTrustSeal[4], ISO16363[5], DIN31644[6]) are all built around OAIS Reference Model (ISO14721)[7] concepts. These concepts inform a number of mandatory responsibilities that include the provision of active **long-term preservation** sufficient to ensure that digital objects (data and metadata) become and remain **independently understandable** to a **designated community** of users that have a defined **knowledge base**.[8] Repository preservation policies, procedures and actions are defined in light of both cultural and technological change.

The FAIR Principles of making data findable, accessible, interoperable and reusable[9] are at the heart of the repository mission. But the Principles themselves do not take account of inevitable changes to the data environment and the needs of data users. The added value of a trustworthy digital repository is the key role they play in enabling data to become and remain FAIR over time.

Any digital system that holds (accepts, stores and provides access to) data that others want can be termed a 'digital repository' in the broadest sense. All repositories exist as an organisation in their own right or are part of a larger organisation. Only an organisation that meets clear criteria, including the provision of active long-term preservation measures for a designated community, can be termed a "trustworthy digital repository". As indicators and tests emerge for the assessment of data and metadata as FAIR (RDA FAIR Data Maturity Working Group[10]) it will become possible to identify whether a TDR is also enabling FAIR data.

Selected extracts from the OAIS Reference Model are provided in Appendix A for context. Appendix B contains a number of selected OAIS definitions.

[4] https://doi.org/10.5281/zenodo.3638211

[5] https://public.ccsds.org/pubs/652x0m1.pdf

[6] https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/siegel.html

[7] https://public.ccsds.org/pubs/650x0m2.pdf

[8] Terms in bold are defined in Appendix B

[9] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

[10] https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines-0

Introduction

Data, whether personal, cultural, scientific or commercial, need to be managed as an asset if we are to maximise their value. The sheer quantity of data available, the tools needed to analyse it, our ability to understand, store and access it all present challenges for our immediate use of data and for ensuring that it remains accessible and reusable in the future.

By data, we mean the original data points of interest (as created and/or collected) plus the metadata that we need to describe, manage and use them.[11] So to be inclusive (and to avoid any discussion of data as singular or plural) we'll refer to this (meta) data as digital objects.[12] We will also acknowledge that some important metadata may be managed 'outside' the digital object. Some core concepts and requirements apply across all digital objects and the people and organisations that manage them, while other ideas and expectations are more specialised, but no less important.

One of the many nodes in the lifecycle of a digital object is the 'repository', and the idea of the Trustworthy Digital Repository (TDR) is seen as a key way of increasing confidence in the interchange of digital objects. The requirements for a TDR cover organisational and technical infrastructure as well as the ability to manage digital objects to agreed standards. The CoreTrustSeal Requirements seek to define and assess an applicant's compliance with these 'core' characteristics. But the requirements also have relevance to the wider range of data services that hold, curate and use digital objects through the lifecycle. The 'repository' or 'service' may be a single dedicated 'organisation', a set of functions within a wider organisation, or a defined partnership of some kind.

In the realm of scientific digital objects and the infrastructures that support them one key factor is the call for FAIR data and this friendly acronym and the ongoing work to define what it should mean in the real world has had enormous impact. Among the areas of adoption is the work to integrate scientific data infrastructure into the European Open Science Cloud (EOSC).[13] But the EOSC itself is indicative of a global trend towards more integrated and interoperable data environments, systems and cultures. The work to apply the FAIR principles to the data ecosystem in a practical way has highlighted the fact that digital objects are inextricably dependent on their context, the infrastructure of people, processes and technology that care for them. FAIR-enabling repositories help to ensure FAIR objects remain FAIR over time.

Work to identify standard indicators for the FAIRness of digital objects[14] and to assess the ability of repositories and other data services to enable FAIRness[15] have confirmed two key assumptions:

1. Existing (meta) data and (meta) data environments are diverse and heterogeneous often even within a single repository or a single digital object.
    - This indicates that reaching FAIRness and Trustworthiness is a journey and that setting standards and pass/fail criteria is not sufficient. We need community and cooperation. This is easy to write, but hard to deliver.
2. We must understand the needs of (future) data users that drive change to digital objects and to repository data services.
    - This is harder to clarify and as hard to deliver, but is equally critical to reaching the goal of digital objects that are curated for FAIRness over time.

The FAIR Principles do not specify how digital objects are made FAIR or for how long they should be kept FAIR. TDR requirements provide this perspective by defining the expectations for long-term digital object preservation. They also clarify that preservation should focus on serving a designated community.

The exact curation and preservation steps required depend on the individual circumstances and expectations.

[11] Software may be the 'data' itself, or be a key component in representing the data.

[12] OAIS differentiates between submission (SIP), archival (AIP) and dissemination (DIP) 'information packages'.

[13] https://www.eosc-portal.eu/

[14] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

[15] https://www.fairsfair.eu/fair-certification

Example 1: If I deposited my French Francs in a bank in 2001 I would expect to have access to the Euro equivalent plus interest in 2021; I would be disappointed to simply receive my original notes and coins. If I loan my collection of ancient Roman coins to a museum I expect them to receive expert care and be returned when agreed; I would be disappointed to be offered an equivalent value in Euros ten years later. What is important is that the consideration has been given to the desirable results and that the progress and outcomes are clear and monitored.

Example 2: If I find a 30-year old research dataset in a data repository, I would expect to have access to this data in a format that I can use now; I would be disappointed if I get the data file in an obsolete format that can't be opened or if the metadata is not sufficient for me to understand the dataset and its origins, and to be certain that the content is authentic.

Preservation for FAIR must go beyond these examples. To enable maximum reuse it is not sufficient to assume that digital objects will be used by people similar to the original researcher and for similar purposes.

The authors of this paper have extensive experience with standards development, assessment, digital object management, providing direct repository support and informing data infrastructure policy. But a functioning data ecosystem, with increasingly complex partnerships and demands for interoperability extends beyond a repository audience with direct experience of archival and repository best practices. Across the vast range of actors in the data space, the two concepts that are both critical to progress and most often misinterpreted are *preservation* and the *designated community*. We'll start with the latter.

Designated Community

We're all doing something for someone. But to deliver an effective service at scale we need a sense of the types of users we have, and how we can meet their needs, also in the future. What do we need to supply to meet the demand? A repository needs a conversation with users to identify what digital objects they want and how they want to use them.

We all hope that our data will interest governments, citizens and journalists as well as scientists. Many repositories have a wide and ever-changing range of users who happen to discover, access and use the data they hold. But when repositories are working to curate data to enable reuse they need to focus their resources on a clear subset of the wider group of actual and potential users. Repositories need to identify their current primary users and their characteristics in order to establish an effective preservation plan that meets the requirements of these users. If repositories identify these users who are their priority for using data then they are identifying one or more communities to be served: this is their designated community. The composition and needs of the designated community may change over time through natural progression or through changes to the repository scope or mission.

It can be easy for a repository to say "*we serve the general public*", but does this suggest that digital objects are findable, accessible, interoperable and (crucially) reusable by a ten year old, by someone without basic computer skills, or even someone who doesn't have a degree in your specialist subject? If not then you've already started to specify the 'knowledge base' of your designated community by making (tacit) assumptions about its members' knowledge, abilities, and competencies, or access to technology. This does not reduce the opportunities to serve wider communities of users and stakeholders. It simply sets a clear focus on what guides acquisitions decisions, preservation actions, and the assessment of TDR status.

A range of interactions with the broader community of users, alongside a clear idea of the designated community can be used to develop periodic evaluation of needs and evolving expectations. This 'community watch' is a critical component of offering a preservation service, but has other wider organisational benefits in terms of user engagement and business intelligence.

Preservation

The level of speciality of the defined designated community (discipline, domain) also informs the preservation actions (changes to data and metadata) to be taken over time. Changes include updates to data formats, to usage licenses, to metadata schema and content (including ontologies) and also include emulation approaches. To achieve TDR status it is necessary but not sufficient to provide technical solutions for deposit, storage and access. A trustworthy digital repository must demonstrate engagement and expertise to identify the needs of the designated community, and provide evidence that it responds appropriately to meet those changing needs.

Direct communication with the designated community of users is not the only route to serving their needs. Repository knowledge and access to external expertise is critical in delivering reusable digital objects now and ensuring their continued value in the future. This includes an understanding of the ethical, legislative and rights conditions as well as 'technology watch' measures to ensure that data and metadata standards and format risks (changes in availability or demand) are addressed through transformation[16] or emulation[17]. To render and interact with digital objects for reuse there may be a wide range of dependent representation information including documentation, software and hardware.

If we've identified "who for" the next question is not simply "how long for". "Forever" is an unkeepable (and untestable) promise. Retention is not the same as preservation. Minimum retention periods provide useful assurances to data funders, depositors and users, as long as they are provided with sufficient information about the level of curation and preservation in place. "For a minimum period after last access or use" is perhaps more flexible, as long as it's clear what decisions and actions may be taken after that time. It's perfectly reasonable to appraise and re-appraise the value of data over time and to decide that it is not worth preserving, or even retaining. But the appraisal process, decisions and consequences should be transparent, documented and managed. The appraisal process is also required to identify the significant properties or essential aspects of the digital objects that need to be preserved in the long term.

Without some level of appraisal we are more likely to overlook the potential long-term value of data and therefore less likely to take early actions to support that value. Long-term preservation requires resources; digital objects will not stay accessible without preservation activities. The cost may increase with a greater level of speciality, but it might also be inflated in a generalist repository with a very broad range of content to preserve. Complexity of data as well as dynamic changes to your designated community, technology and legal framework will also influence cost. In many cases, acquisition and ingest (curation) are the most costly preservation phases but economies of scale are possible.[18] But in all cases making relatively small investments in digital object management throughout the lifecycle can make any eventual decision to preserve less costly.

Once the informed decision has been made to 'not delete' data the appropriate actions to be taken must be identified. If the minimum period of retention is measured in years then some change to the needs of the designated community, or the technical supporting infrastructure is to be expected. Many of the measures we take to reduce risks to digital objects in long-term preservation are identical to the measures necessary to cope with the 'next round of change' that even a short-term data curator might face. Trustworthy digital repositories are designed to keep addressing the next round of change for the long term.

All organisations that hold digital objects face two key challenges in ensuring that digital objects are authentic: avoiding unintended change (data integrity) and managing intended change (to hardware, software, media or objects through copies, versions, change logs etc.). These inform provenance and therefore a basis for trust. Both depend on a degree of information security, authentication, authorisation etc. The technical management of these relatively static 'bits' is sometimes referred to as 'bit-level preservation'. Multi-copy, multi-media, multi-location and integrity-assured bit-level storage measures to address the long tail of 'at risk' data are necessary, but not sufficient, for a trustworthy digital repository.

Preservation Tiers

Any repository must offer basic data and metadata deposit, storage and access functionality. In addition to maintaining the systems that store the bits they will need to update their deposit metadata schema or amend their resource discovery and access metadata schema and interfaces over time.

A technical service provider focussed on data storage and access could forward migrate metadata to new versions, or even map and move to new schemas over time. It might feasibly provide the automated migration of data to new formats if old formats

[16] https://en.wikipedia.org/wiki/Data_transformation

[17] https://en.wikipedia.org/wiki/Emulator

[18] Beagrie, N and Lavoie, B and Woollard, M (2010). Keeping Research Data Safe 2: Final Report. Joint Information Systems Committee (JISC). Available at http://repository.essex.ac.uk/id/eprint/2147

become obsolete. By addressing the risk of metadata schema and file format obsolescence, these steps take the preservation of digital objects beyond merely storing the bits, but they will still fall short of the expectations of a trustworthy digital repository if they do not, in addition, take steps to maintain data and metadata reusability for their defined designated community in accordance with this community's (evolving) needs. This entails ensuring that data and metadata are - in OAIS terms - "independently understandable" to the designated community, and that potential losses of functionality that may occur when data are converted to new formats are carefully evaluated in light of the designated community's data practices. Thus, the critical changes as we move up the tiers from bit-level preservation are around rights and responsibility. A trustworthy digital repository must have the rights (as well as the knowledge and the skills) to decide what preservation actions need to be taken on data and metadata over time. This includes responding to changes in demand for data formats and metadata schemas from the designated community.

The level of outsourcing permitted as a TDR can vary, but it is clear that the decisions on preservation actions must remain with a trustworthy digital repository even if the actions are carried out by others. It is not sufficient to devolve decisions about preservation actions to the original data producers or depositors; those experts may not be available for as long as the data is preserved. As noted in the OAIS Reference Model the "*Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information*".

Preservation in specialist and generalist repositories

Some digital objects will be cared for by generalist repositories with a fairly broad-based designated community. For these repositories it may be sufficient to offer metadata for general purpose resource discovery and use (e.g. via google search, or using metadata schemas like DataCite or Dublin Core). For a specialist repository with a defined discipline or domain of expertise there is likely to be a narrower designated community (i.e. an assumption of greater knowledge in the speciality) and a concurrent expectation that the metadata for resource discovery and description are fit for these more specialised needs (e.g. DDI for social science digital objects, CMDI for language resources, CIDOC CRM for cultural heritage or IVOA Technical Specifications for astronomy).

To become certified as a TDR, a specialist repository might accept a more limited and specialised set of file formats for deposit to ensure they can be preserved and made accessible. Conversely, a more generalist repository with no restrictions on deposited file formats will be expected to demonstrate that it can preserve this wider range over time.

But in general the expectations of a disciplinary or domain repository are that it delivers all of the functions and services of a generalist repository (for data meeting its appraisal and selection criteria) plus the additional functions required by its specialist designated community. These distinctions between a generalist trustworthy data repository and one which serves a specialist designated community become important in the face of recommendations that researchers deposit their data in disciplinary trustworthy digital repositories wherever possible[19],[20] .

Preservation ensures that data users can find stored digital objects, access them, and interact with them in ways that, with the help of associated metadata, software and hardware, allows them to understand and use them. Preservation further ensures that this remains the case if (and when) the data, storage, access, location, representation or descriptive information, or the needs and knowledge base of the designated community, change over time. This cannot be assured by relying on original data producers.

The level to which a repository can provide FAIR digital objects and meet the needs of its designated community initially depends on the level of curation. The level to which it can meet those needs and retain the FAIRness of digital objects over time depends on preservation. Not all digital objects require the same degree of care. Providing any level of curation and preservation has costs. Failing to meet the required level of care has a lower cost but a higher likelihood of a negative outcome (risk of losing data, or more likely losing the understandability and re-usability of data).

It must be clear to data funders, depositors and users which types of preservation a repository is capable of and what level of curation and preservation it offers for a particular digital object or collection of objects. It must be clear to all actors when the level of preservation changes and why, and ideally it should be possible for other repositories to take over the data to provide a higher level of preservation.

[19] Science Europe Guide to the Practical Alignment of Research Data Management

[20] NIH, Selecting a Repository for Data Resulting from NIH-Supported Research

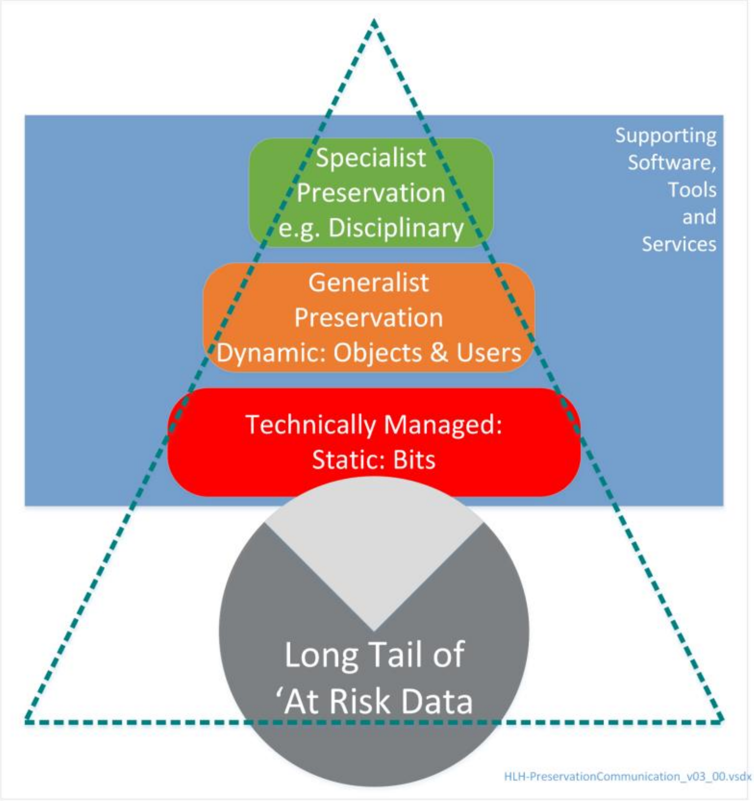

Diagram 1: Communicating Preservation: Technical, Generalist and Specialist tiers

In diagram 1 above, we present the tiers of technical, generalist and specialist measures that a data-holding organisation can provide. As the diagram progresses through the co-dependent tiers of storage, curation and preservation risk is reduced. These may depend on a range of supporting software, tools and services but the repository maintains final responsibility. Data that is technically managed reduces the long tail of at-risk data. This is necessary, but not sufficient to be a TDR. A repository must actively preserve to meet the needs of a generalist or specialist designated community.

Conclusion

Enabling FAIR data and achieving Trustworthy Digital Repository status are a journey that requires expertise, investment and cooperation. No one is expecting preservation personnel to have perfect psychic foresight, or infinite resources to address all possible futures. The standards and best practices are designed to guide the delivery of quality data services, now and in the future, through organisational infrastructure, digital object management and technical and security measures.

Many actors and multiple tiers of data storage, protection and preservation are needed to address the long tail of at-risk data and metadata. When depositing my currency in a bank or my gold coins in a museum it is vital that I can clearly identify the parties I am trusting, that we mutually understand what has been put into their care, and what the expected and permitted actions and outcomes of that 'curation' will be. What is important is that consideration has been given to the desirable results and that the progress and outcomes are clear and monitored. The co-dependant levels of storage, curation and preservation, as well as the temporal and specialist or generalist nature of the care provided, are key factors that must be transparent to all stakeholders.

The evolving FAIR indicators for digital objects will provide a 'snapshot' of current FAIRness. It remains to be defined how these FAIR characteristics can be enabled and preserved over time by the organisations that care for them. Some of these organisations are candidates for trustworthy digital repository status via standards like the CoreTrustSeal because they can demonstrate that they take active preservation steps in line with the needs of a defined designated community. But other types of organisation also play important roles in the research data lifecycle, and their impact on FAIRness over time must be evaluated.

Appendix A: Brief OAIS Context: Preservation and Designated Community

For a more detailed overview see OCLC Research, DPC Technology Watch Report The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)[21] (Lavoie, 2014).

The OAIS Model defined "*what is required for an archive to provide permanent, or indefinite Long Term, preservation of digital information*" (OAIS, 2012). An OAIS is "a*n Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community*" where "*The information being maintained has been deemed to need Long Term Preservation, even if the OAIS itself is not permanent*". "L*ong Term digital information preservation and access*" is "*is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community*".

The mandatory responsibilities of an OAIS are to

- Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.
- Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.
- Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity

'Temporary' organisations or data initially assumed to be of 'temporary' value are also in scope because though "*some facilities holding information may themselves be temporary, some or all of their information may need to be preserved indefinitely*" and "*when taking into consideration the rapid pace of technology changes or possible changes in a Designated Community, there is the likelihood that facilities, thought to be holding information on a temporary basis, will in fact find that some or much of their information holdings will need Long Term Preservation*"

Appendix B: Selected OAIS Reference Model Definitions

Archive: An organization that intends to preserve information for access and use by a Designated Community.

Designated Community: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time.

Digital Migration: The transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by three attributes: – a focus on the preservation of the full information content that needs preservation; – a perspective that the new archival implementation of the information is a replacement for the old; and – an understanding that full control and responsibility over all aspects of the transfer resides with the OAIS.

Independently Understandable: A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.

[21] http://dx.doi.org/10.7207/twr14-02

Knowledge Base: A set of information, incorporated by a person or system, that allows that person or system to understand received information.

Long Term: A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future.

Long Term Preservation: The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term

Open Archival Information System (OAIS): An Archive, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. It meets a set of responsibilities, as defined in section 4, that allows an OAIS Archive to be distinguished from other uses of the term ‘Archive’. The term ‘Open’ in OAIS is used to imply that this Recommendation and future related Recommendations and standards are developed in open forums, and it does not imply that access to the Archive is unrestricted.

Transformation: A Digital Migration in which there is an alteration to the Content Information or PDI of an Archival Information Package. For example, changing ASCII codes to UNICODE in a text document being preserved is a Transformation

Authors and Affiliation

Author Name	Project Affiliation	Organisational Affiliation	Author Identifier
Hervé L’Hours	SSHOC, FAIRsFAIR	UK Data Archive, UK Data Service, University of Essex	0000-0001-5137-3032
Mari Kleemola	EOSC Nordic, SSHOC	Finnish Social Science Data Archive, Tampere University	0000-0001-8855-5075
Ilona von Stein	FAIRsFAIR	Data Archiving and Networked Services	0000-0003-3179-0773
René van Horik	FAIRsFAIR, SSHOC	Data Archiving and Networked Services	0000-0001-6899-760X
Patricia Herterich	FAIRsFAIR	Digital Curation Centre (DCC)	0000-0002-4542-9906
Joy Davidson	FAIRsFAIR	Digital Curation Centre (DCC)	0000-0003-3484-7675
Olivier Rouchon	FAIRsFAIR	Centre Informatique National de l’Enseignement Supérieur (CINES)	0000-0002-1816-5546
Mustapha Mokrane	FAIRsFAIR	Data Archiving and Networked Services	0000-0002-0925-7983
Robert Huber	FAIRsFAIR	MARUM - Center for Marine Environmental Sciences, University of Bremen	0000-0003-1648-2123