# Data Curation and Preservation

Keith Jeffery(✉)

Keith G Jeffery Consultants, 71 Gilligans Way, Faringdon SN7 7FX, UK
keith.jeffery@keithgjefferyconsultants.co.uk

**Abstract.** Data is a valuable resource. In some scientific disciplines, experiments can be redone to reproduce the data. In environmental sciences, the observations and measurements of the earth and its surroundings commonly can be made only once: each time point records uniquely the state of the many earth processes. This demands that environmental data - structured to information - is preserved in such a way that it may be reused. Phenomena like the ozone hole, biodiversity and climate change depend on data curated over a long period of time. However, it is not just the data that must be curated. The software used to process and analyse the data - or more accurately an executable specification of the software - must be preserved along with associated libraries and computing operational environment. Information on the equipment and sensors used must be preserved since this affects the relevance and quality of the data for future use. Equally challenging is the decision to discard data - for reasons of costs of storage (although that is reducing rapidly) or cost of curation. Curation is blended inextricably with cataloguing and provenance and the core requirement is for rich metadata to characterise the digital asset for all three purposes.

**Keywords:** Data · Information · Preservation · Curation · Storage · Metadata · Cataloguing, provenance

## 1 Introduction, Context and Scope

"*Digital curation is the selection, preservation, maintenance, collection and archiving of digital assets. Digital curation establishes, maintains and adds value to repositories of digital data for present and future use. This is often accomplished by archivists, librarians, scientists, historians, and scholars*" (Wikipedia)[1].

Cataloguing, Curation and Provenance are commonly grouped together since the metadata, workflow, processes and legal issues associated with each have a high degree of intersection in recorded metadata attribute values and therefore rather than generating independent systems a common approach is preferable. Moreover, there are strong interdependencies with identification and citation, with AAAI (Authentication, Authorisation, Accounting Infrastructure), with processing, with optimisation, with modelling and with architecture.

[1] https://en.wikipedia.org/wiki/Digital_curation.

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research
Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 123–139, 2020.
https://doi.org/10.1007/978-3-030-52829-4_7

A key aspect of curation is the interplay between governance and technology. Finding technological solutions to satisfy the principles of governance is not always easy. The increased acceptance of the Data Curation Lifecycle and the increasing use of Data Management Plans (DMPs) evidences this. Another key aspect is involving the researchers in the decision making of what to keep and what to discard; this provides motivation for the process of curation, including the provision of appropriate metadata.

## 2 Curation Within ENVRIplus

The ENVRI community observes and analyses many aspects of Earth's changing phenomena. Observations and analyses today may be needed or reviewed in ways that are impossible to predict. Consequently, preparing the platform for future researchers as best we can by investing in curation has to be a key element of the ENVRI research culture with broad support by Research Infrastructures (RIs) and researchers. This requires leadership, education and collaborative development.

The ideal curation culture will ensure – via an appropriate IT system including both technological and governance aspects - the availability of digital assets through media migration to ensure physical readability, redundant copies to ensure availability, appropriate security and privacy measures to ensure reliability and appropriate metadata to allow discovery, contextualisation (for relevance and quality) and use, including information on provenance and rights.

At the curation stage of the lifecycle we record metadata concerning quality. Such metadata is – by its nature – domain specific and to some extent subjective. The required quality of the asset described by the metadata depends heavily on the purpose to which it is to be put. Decisions that are of broad scope and/or urgent may require only summary quality metadata whereas decisions relating to critical and detailed information such as in reproducibility of research may need detailed technical quantitative parameters recorded in the metadata. Thus, the end-user has to decide – based on the metadata available, guidelines established by governance and training to develop the skills – whether the asset is of appropriate quality for the intended purpose and whether – based on cost-benefit analysis - it should be curated. Clearly, the richer and more comprehensive the metadata providing context, the better judgement on quality can be made. The quality processes for some RIs in the environmental sciences have been studied in [1] and both a quality taxonomy and potential improvements recommended.

There has been significant progress over the period of the ENVRplus project: (1) the RIs appreciate the curation lifecycle; (2) the RIs generally have developed DMPs usually using the DCC (Digital Curation Centre) template appropriate for H2020 (EC Horizon 2020) projects; (3) the RIs appreciate the interplay between curation and both cataloguing and provenance; (4) the RIs understand the requirements for rich metadata to effect curation (and also cataloguing and provenance); (5) some RIs are planning future evolution utilising these principles.

## 3 Current Curation Activity

In the ENVRI community, there is a curation activity [12, 13]. Starting from a relatively low base at the beginning of the ENVRIplus project, curation activity has risen steadily encouraged by the presentations at ENVRI meetings and by the collection of information on curation associated with requirements collection.

### 3.1 Curation Lifecycle

The desirable lifecycle is represented by a DCC diagram, as shown in Fig. 1. The DCC in the UK is responsible for advising researchers and others on digital curation. The lifecycle model emphasises the steps in curation, the information required and the decisions to be taken at each step.

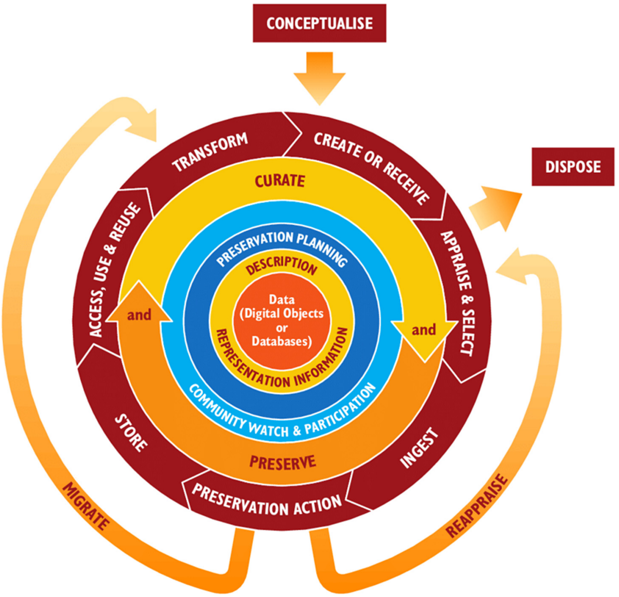

**Fig. 1.** The curation lifecycle model ("The DCC Curation Lifecycle Model", JISC/DCC, http://dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf) from DCC.

### 3.2 Data Management Plan

A Data Management Plan is defined as "A data management plan or DMP is a formal document that outlines how you will handle your data both during your research, and after the project is completed" (Wikipedia)[2].

The ENVRIplus RIs now generally have DMPs and utilise these as a basis for internal policymaking, road mapping, technological planning and governance of asset management, the latter within the framework of governance established by the RI e.g. the governance of a consortium through a consortium agreement. They are also used in the context of external agreements on transnational access.

[2] https://en.wikipedia.org/wiki/Data_management_plan.

### 3.3 OAIS Reference Model

The Open Archival Information Systems (OAIS) Reference Model - ISO 14721:2002 [2] - provides a generic conceptual framework for building a complete archival repository and identifies the responsibilities and interactions of Producers, Consumers and Managers of both paper and digital records. The standard defines the processes required for effective long-term preservation and access to information objects while establishing a common language to describe these. It does not specify an implementation but provides a framework to make a successful implementation possible, through describing the basic functionality required for a preservation archive. It identifies mandatory responsibilities, and provides standardised methods to describe a repository's functionality by providing detailed models of archival information and archival functions [3]. Some RIs have considered OAIS as a framework but none has implemented it fully, although the concepts of Submission Information Package (SIP): which is the information sent from the producer to the archive, Archival Information Package (AIP): which is the information stored by the archive and Dissemination Information Package (DIP): which is the information sent to a user when requested have influenced the curation work in environmental and Earth science RIs.

In order to populate such a framework, a rich metadata element set is required. Much work has been done investigating various metadata standards to assess their suitability for curation (as well as for cataloguing and provenance). Within the work of the RDA (Research Data Alliance) MIG (Metadata Interest Group) – of which the chapter author is co-chair – a set of metadata elements in a structure for the purposes of curation, cataloguing and provenance according to the FAIR principles[3] has been proposed[4].

### 3.4 RDA (Research Data Alliance)

The Research Data Alliance has groups working on curation, provenance and catalogue metadata as well as citation. Clearly there is a benefit to ENVRIplus in alignment with the evolving RDA metadata recommendations which assist greatly not only in curation but also cataloguing, provenance and citation leading to improved discovery, contextualisation (for relevance and quality), interoperability, scientific reproducibility, and general governance of research assets. However, the RDA work is brought together with that of other groups in the specification of metadata[5]. The other groups are either domain-specific (e.g. in agriculture) or cross-cutting (e.g. in citation).

RDA proposed some metadata principles which are now generally accepted in that community:

- The only difference between metadata and data is a mode of use;
- Metadata is not just for data, it is also for users, software services, computing resources;
- Metadata is not just for description and discovery; it is also for contextualisation (e.g. relevance, quality, restrictions (rights and costs)) and for coupling users, software and computing resources to data (to provide a VRE);

[3] https://www.force11.org/group/fairgroup/fairprinciples.

[4] https://drive.google.com/drive/folders/0B8FnM3PsoL2dd2RnYVBmcjRMYXc.

[5] https://www.rd-alliance.org/groups/metadata-ig.html.

- Metadata must be machine-understandable as well as human-understandable for autonomicity (formalism);
- Management (meta)data is also relevant (e.g. research proposal, funding, project information, research outputs, outcomes and impact); and furthermore, a metadata element set that covers all the uses of metadata (not just curation):

  - Unique Identifier (for later use including citation);
  - Location (URL);
  - Description;
  - Keywords (terms);
  - Temporal coordinates;
  - Spatial coordinates;
  - Originator (organisation(s)/person(s));
  - Project;
  - Facility /equipment;
  - Quality; Availability (licence and persistence) including curation duration;
  - Provenance;
  - Citations;
  - Related publications (white or grey);
  - Related software;
  - Schema;
  - Medium/format;

It should be noted that many elements within this set have an internal structure (syntax) and semantics (meaning – usually represented by an ontological structure with term explanation and relationships) and so are not simple attributes with values. The RDA groups continue working on 'unpacking' the elements to a form suitable for discovery, contextualisation and action by both humans and computers.

# 4 Problems to Be Overcome for Curation in ENVRI

## 4.1 Current State

Some important problems associated with curation were discovered during requirements collection:

> *"ENVRI research communities will expect an integrated and seamless curation service that supports their routine work well and that opens paths for innovative research. This will require engagement from the practising domain scientists to help the ICT experts deliver relevant curation systems"* [4].

The incremental progress achieved for each problem is documented below:

### Motivation

*Problem to Be Overcome:* There is little motivation for researchers to curate their digital assets. At present curation activity obtains no 'reward' such as career preferment based on data citations. In some organisations curation of digital assets is regarded as a librarian function but without the detailed knowledge of the researcher the associated metadata is likely to be substandard. Increasingly funding agencies are demanding curation of digital assets produced by publicly funded research.

*Progress Achieved:* Motivation has increased significantly – but not sufficiently yet. Use cases that provided significant scientific results dependent on curation are well-known and have provided motivation. The requirement by funding agencies for DMPs has also caused increased interest in and compliance with curation principles. Finally, the increasing availability to researchers of 'data stewards' and curators are improving curation.

**Business Model**

*Problem to Be Overcome:* Curation involves deciding what assets to curate and of those, for how long they should be kept. Determining an appropriate duration of retention for a digital asset is a problem; economics and business models do not manage well the concept of infinite time. First a business justification is needed in that (a) the asset cannot be collected again (i.e. it is a unique observation, experiment); (b) the cost of collecting again (by the same or another researcher) is greater than the cost of curation.

*Progress Achieved:* Awareness of the data curation lifecycle (within the research lifecycle) has increased leading to better governance and improved curation decisions.

The economic problem remains but decreasing costs of both storage and processing argue for increased curation by improving the cost/benefit ratio.

The major cost of curation is in expert staff providing guidelines and protocols and also – ideally – associated software tools. Increasing automation and autonomicity of curation processes will further reduce costs leading to an acceptable economic model in time.

**Metadata**

*Problem to Be Overcome:* Metadata collection is expensive unless it is automated or at least partially automated during the data lifecycle by re-using information already collected. Commonly, metadata is generated separately for discovery, contextualisation, curation and provenance when much of the metadata content is shared across these functions. A comprehensive but incrementally completed metadata element set is required that covers the required functions of the lifecycle. It needs sufficient application domain data that other specialists in that domain will be able to find and correctly interpret the associated data. Making the metadata handling facilities and tools that use them, such as workflows and data management, available to practical researchers to help them in their daily work, encourages them to invest in metadata, improves the quality of domain metadata and therefore facilitates the later curation processes [5].

*Progress Achieved:* Awareness of the need for - and benefits to be derived from – rich metadata is increasing substantially in the RIs as they evolve. This evolution is driven by researcher aspirations and requirements and is supported by improving technology. At present there are many metadata standards - international (e.g. ISO19115), local variants of international (e.g. INSPIRE[6] or APs (Application profiles) of DCAT[7], OpenSearch geoxtension[8]) and community/local all used in RIs within ENVRI. However, interconversion among all n of them requires n(n − 1) converters. Using a common canonical rich metadata schema as the 'switchboard' for interoperation between RIs reduces this to n convertors.

The co-development of rich metadata cataloguing, curation and provenance is a journey taking the RIs from a processing and governance environment where much human effort is required to re-use the assets with poor metadata to an automated environment with much re-use of the assets leveraged by rich metadata.

The cost of metadata creation is high. However, increasingly it is collected incrementally along the research workflow so reducing the perceived cost at each step. With rich metadata used for cataloguing, curation and provenance functions the scientific benefit increases relative to the costs of collection.

The utilisation of CERIF[9] additionally to CKAN[10] as the metadata standard for interoperation in ENVRIplus will improve the situation even further because of its much richer syntax and semantics (providing a superset canonical standard for interoperation) and its provision of referential and functional integrity.

### Process

*Problem to Be Overcome:* The lifecycle of digital research entities is well understood and it needs process support. The incremental metadata collection aspect is critically important for success. Workflow models – if adapted to such an incremental metadata collection with appropriate validation –are likely to be valuable here [6].

*Progress Achieved:* Within some RIs we see increasingly the use of workflows (and, indeed, in some, automation of workflow deployment across multi-cloud or multiple processing environments managed by rich metadata). This allows for incremental metadata collection as predicted (with consequent benefits) but also highlights the need for rich metadata if automated processing – and thus reduction of human costs in research - is to be achieved. This was demonstrated in the PaaSage project[11] where the chapter author was scientific coordinator.

### Curation of Data

*Problem to Be Overcome:* It may be considered that curation of data is straightforward –but it is not. First the dataset may not be static (by analogy with a type-specimen in a museum); both streamed data and updateable databases are dynamic thus leaving management decisions to be made on frequency of curation and management of versions with obvious links to provenance. Issues related to security and privacy change with

[6] https://inspire.ec.europa.eu/metadata/6541.

[7] https://www.w3.org/TR/vocab-dcat-2/.

[8] https://www.opengeospatial.org/standards/opensearchgeo.

[9] https://www.eurocris.org/cerif/main-features-cerif.

[10] https://ckan.org/portfolio/metadata/.

[11] https://paasage.ercim.eu/.

time and the various licences for data use each have different complexities. The data may change ownership or stewardship. Copies may be made and distributed to ensure availability but then have to be managed in systems such as LOCKSS[12]. Derivatives may be generated and require management including relationships with the original dataset and all its attendant metadata.

*Progress Achieved:* After the first half of the project, the RIs have increased their awareness – and appreciation – of this problem.

The relationship with provenance and cataloguing is clear – and the need for an integrated rich metadata catalog to cover all these processing and governance requirements in an integrated and consistent fashion is also becoming clear to the RIs.

The need for metadata covering not only description of the asset and its history, but also the persons and organisations - backed by funding – responsible is now understood.

Technology for the management of distributed copies – and their partitioning/replication/migration for processing efficiency overcoming latency – in a multi-cloud environment is being developed in the MELODIC project[13] where the chapter author is a consultant to the project.

The RDA Data Citation Working Group[14] has produced a recommendation for managing citation to parts of and versions of datasets. This relies on appropriate curation of the versions and services to define the partition.

### Curation of Software

*Problem to Be Overcome:* Software written 50 years ago, is unlikely to compile (let alone compose with software libraries and execute) today. Indeed, many items of software, such as the workflows behind a scientific method, will either not run or give different results, six months later. Since many research propositions are based on the combination of the software (algorithm) and dataset(s) then the preservation and curation of the software become very important. It is likely that in future it will be necessary to curate not only the software but also a specification of the software in a canonical representation so that the same software process or algorithm can be reconstructed (and ideally generated) from the specification. This leaves the question of whether associated software libraries are considered part of the software to be curated or part of the operating environment (see below). Very often software contains many years-worth of intellectual investment by collaborating experts. It is not unusual for the software to encode the 'scientific method' used by the researcher which may be less well (or less formally) documented elsewhere (e.g. scholarly publications). This makes software very valuable and hard to replace. Taking good care of such assets will be a requirement for most research communities.

*Progress Achieved:* The issue was novel to most RIs when introduced in ENVRIplus Task T8.1 and recorded in Deliverable D8.1 [7]. The requirement is now appreciated but the metadata systems in use in most RIs are incapable of providing a technological

[12] https://www.lockss.org/.

[13] http://melodic.cloud/.

[14] https://www.rd-alliance.org/groups/data-citation-wg.html.

solution. It is further complicated because many developers – including those in some RIs – use GitHub[15] and related (or similar) technology to manage software development including versions, copies, compositions and deployments.

There is – as yet – no generally accepted way of managing this from both the technological and governance points of view. From an ENVRIplus perspective, the best we can do is to use rich metadata to catalogue the software and its evolution and monitor work elsewhere that will provide appropriate solutions.

### Curation of Operational Environments

*Problem to Be Overcome:* It is necessary to record the operational environment of the software and dataset(s). The hardware used – whether instrumentation for collection or computation devices – has characteristics relating to the accuracy, precision, operational speed, capacity and many more. The operating system has defined characteristics and includes device drivers – i.e., a software library used by the application. It is a moot point whether software libraries belong to the application software or to the operational environment for the purposes of curation. Finally, the management ethos of the operational environment normally represented as policies requires curation.

*Progress Achieved:* The issue was novel to most RIs when introduced. The requirement is now appreciated but the metadata systems in use in most RIs are incapable of providing a technological solution.

There appears to be no generally accepted solution available. The best we can do in ENVRIplus is to collect rich metadata covering the operational environments and monitor external developments to find solutions as they are developed.

Increasingly, there appears to be a partial solution in containerisation using e.g. Docker[16] or Kubernetes[17]. Unlike Virtual Machines (VMs) (which have the contents of the container plus the operating system and are thus heavier on resources) containers include just the application and associated libraries and runtime environment and thus can be moved from one operating system environment to another, utilising the operating system kernel read-only and permitting writing to the container through its own 'mount' (access to the container).

### Curation of 'Raw' Data Collected by Sensors or Instruments

*Problem to Be Overcome:* This is a special class of operational environment of importance to RIs in environmental science. The problems are manifold due to the data collection volume, velocity, variety, veracity and value and the difficulties of analytics, simulation and visualisation of streamed data.

*Progress Achieved:* While the requirements collected early in the project concentrated on the curation of validated or part-processed data, some RIs require curation of (at least some) raw data to allow subsequent reprocessing in calibration for precision and accuracy. Some examples illustrating the variety of practice are given below. EMSO

[15] https://github.com/.

[16] https://www.docker.com/.

[17] https://kubernetes.io/.

has distributed observatories with differing policies. In contrast Euro-Argo centralises quality control and curation. IAGOS validates the raw data manually or automatically before curation. ICOS stores (a kind of curation) raw sensor data collected at the stations and curates validated data. LTER does ingestion and quality control (curation) at individual sites. SeaDataNet relies on local centres curating quality-controlled data. An aspect particularly relevant increasingly to ENVRI communities is semi-automated curation of metadata which can be achieved if instrument metadata is available (SensorML[18] or SSNO[19]) and e.g. linked by PIDs with the incoming data stream[20].

### 4.2 A Longer-Term Horizon

There is some cause for optimism. Work within the ENVRIplus project has increased knowledge and understanding among the RIs and has exposed the issues and challenges to be addressed. A list of reasons for the optimism is:

1. Media costs are decreasing – so more can be preserved for less (and the cost reduction hopefully matches the expansion of volume). Media costs have decreased even more in the last 24 months and the trend shows no sign of changing;
2. Awareness of the need for curation is increasing; partly through policies of funding organisations and partly through increased responsibility of some researchers. The awareness has increased substantially not only through the efforts of ENVRIplus but also international efforts such as RDA and the FAIR initiative. The link with open science (i.e. open access to scholarly publications and datasets) is an effective driver.
3. Research projects in ICT are starting to produce autonomic systems that could be used to assist with curation. In particular MELODIC (mentioned above) is offering solutions combining curation and deployment.
4. Increasing standardisation of metadata and approaches to curation based on rich metadata are emerging and it is to be expected that this will continue producing richer and more effective curation services.

The cost of collecting metadata for curation remains a problem. Reducing storage costs mean that more data (even raw data to allow later re-processing before interpretation) can be stored. However, the major cost is that of creating appropriate metadata for the purposes of curation and subsequent discovery, contextualisation (including provenance) and action on the asset. The relative cost against benefit is reduced considerably by collecting the metadata once and using it for curation, cataloguing and provenance. Incremental collection along the workflow with re-use of existing information has been shown to decrease costs – but particularly to decrease researcher resistance to providing metadata - further. Improving techniques of automated metadata extraction from digital objects offer a further possibility of cost-reduction. There was some hope that they

[18] https://www.opengeospatial.org/standards/sensorml.

[19] https://www.w3.org/TR/vocab-ssn/.

[20] https://www.rd-alliance.org/group/persistent-identification-instruments/case-statement/persistent-identification-instruments.

may reach production status in the ENVRIplus time frame [8]. At present – although progress has been made – there are no appropriate systems although research indicates some cause for optimism.

## 4.3 Issues and Implications

### Lack of Common Metadata Elements

*Issues and Implications:* Commonality of metadata elements across curation, provenance, cataloguing (and more) implies that a common core metadata scheme should be used for interoperability – possibly with extensions for particular domains where interoperability is not required.

*Ongoing Work:* The joint work especially with cataloguing - and following the recommendations from both the ENVRIplus cataloguing activity and the architecture – has led to the development of two catalogues, one using CKAN as in EUDAT B2FIND[21] and the other using CERIF as used in EPOS[22]. Experiments are underway to evaluate the two approaches for capability as the core metadata scheme.

### Metadata Collection Expense

*Issues and Implications:* Metadata collection is expensive so incremental collection along the workflow is required: workflow systems should be evolved to accomplish this and scientific methods and data management working practices should be formalised using such workflows to reduce chores and risks of error as well as to gather the metadata required for curation;

*Ongoing Work:* There is evidence of increased use of workflows in the RIs although many are human-driven and not automated. Nonetheless, this provides the governance process to ensure incremental metadata collection to provide the required rich metadata.

### Automated Metadata Extraction

*Issues and Implications:* Automated metadata extraction from digital objects shows promise but production system readiness is some years away. However, metadata provision from equipment-generated streamed data is available;

*Ongoing Work:* This has been monitored but the current systems are not yet at production status sufficient to be recommended to the RIs

### DCC Recommendations

*Issues and Implications:* ENVRIplus should adopt the DCC recommendations;

*Ongoing Work:* Following acceptance by the RIs, this is achieved. However, implementation is incremental.

[21] https://www.eudat.eu/services/b2find.

[22] https://www.epos-ip.org/.

### RDA Tracking and Involvement

*Issues and Implications:* ENVRIplus should track the relevant RDA groups and – ideally – participate.

*Ongoing Work:* Following acceptance by the RIs both tracking and participation have been pursued actively. Of particular relevance is the work on the RDA Metadata Element set which could be a candidate for a future common metadata scheme.

### Education and Awareness

*Issues and Implications:* ENVRIplus should consider educational and practical steps to increase awareness of curation issues for all practitioners, particularly those concerned with curation organizational and technical strategy – collaboration and coordination could reduce the cost of this.

*Ongoing Work:* Curation has been presented at ENVRI meetings and elsewhere to raise awareness and encourage best practice in both governance and technical solutions.

The appreciation of the data curation lifecycle and the increasing use of DMPs is an achievement. The appreciation of the need for rich metadata for curation (alongside cataloguing and provenance) is also an achievement.

# 5 Architectural Design for Curation in ENVRI

## 5.1 Context

### 5.1.1 Initial State

At the beginning of the project, we asserted three aspects of the then-current state. Each, below, is supplemented by the work done during the project:

1. Technologies are available for curation but they may not be compatible with those for cataloguing and provenance. There has been a rapid and voluminous increase in understanding the need – for technological and governance reasons – to utilise one common metadata standard (in each RI and for interoperation across RIs) covering cataloguing, curation and provenance. Furthermore, it is widely understood and appreciated that this metadata standard has to be rich in syntax and semantics.
2. Governance principles for curation were lacking widely among the ENVRI community. The appreciation of the Data Lifecycle (within the research lifecycle) and the increasing use of DMPs has seen a marked improvement in governance.
3. Most RIs in the ENVRI community appreciate the importance of curation but are not practising it – partly because existing used metadata standards do not support it explicitly and/or can only be made to support it partially. All RIs appreciate the importance of curation and understand the rationale behind the WP8 work towards a rich metadata standard for curation (as well as cataloguing and provenance).

Further work on curation has considered also other, wider, aspects. In particular:

1. The use of personal data;
2. Fixity or preservation of state against possible data corruption.

The use of personal data – even in open science – is a contentious issue. The GDPR[23] (General Data Protection Regulation) of the EU makes provision for protecting personal data and its use. In open science, the name of a person, their institution, the equipment they use, their publications and their research assets are highly relevant to contextualisation (assessing relevance and quality for a new purpose). At present, there is no case law testing the limits of GDPR so this requires tracking and incorporating statements based on any judgements into the governance of RIs and their management (including curation) of data.

Environmental research data is the evidence base for some active political discussions, especially concerning climate change, utilisation of resources and pollution. Clearly, for environmental research, it is essential to have the observations made at a particular location and time preserved (possibly after assessment for accuracy, precision and/or any calibration corrections, smoothing or aggregation). This requires appropriate security to protect the integrity of the research product (asset) against 'tampering'.

### 5.1.2 Current State

It is clear that in the ENVRIplus project timespan the RIs have appreciated the need for a common rich metadata standard covering not only curation but also cataloguing and provenance (chapter 8 and 12). The requirement for protection of personal data and assurance of integrity (including fixity) underlines the need for rich metadata appropriate for enforcing access control. The ICT team has been working towards this and has been evaluating the solutions described in within the context of the requirements and architecture.

The final architectural solution for curation post-ENVRIplus will be decided as a result of that evaluation.

## 5.2 Architectural Design

### 5.2.1 Introduction

The initial design for curation was based not just on the state of the art and requirements for curation, but also for cataloguing and provenance (and also identification, citation and processing) for the reasons outlined above. The design consists of two components: the catalogue metadata and the curation processes. The final design confirms the initial design and adds detail.

### 5.2.2 Catalogue Metadata

The catalogue – for the purposes of curation – needs to describe the asset to be curated with rich metadata. The metadata must provide sufficient information for asset discovery, contextualization (for relevance and quality) and action. This is analogous to – but goes

[23] https://eugdpr.org/.

beyond in the area of action – the FAIR principles. In the case of curation, the action is to ensure an asset can be (a) made available when required; (b) is understandable to human and computer systems. The use of a logic representation provides advantages in deduction (facts from rules) and induction (rules from facts) which reduces potentially the metadata input burden and increases the validity of the metadata. Furthermore, because of versioning and the relationship to provenance, the metadata must include temporal information.

This system design aspect, therefore, depends on the cataloguing activity of ENVRIplus and to some extent on the Provenance activity, all within the overall architectural design.

However, the required metadata elements can be specified, derived from the use cases and requirements and the work of the Metadata Interest Group (and its sub-groups) of RDA (see above under 'State of the Art') which attempts to bring together experience and best practice from many international and national domain-specific efforts at standardising metadata for multiple uses, including curation. The base entities (objects) typically required (but note these may be complex with internal structure (syntax) and semantics) are:

Research Product (i.e. asset), Person, Organisation, Project, Research Publication, Citation, Facility, Equipment, Service, Geographic bounding box, Country, Postal address, Electronic address, Language, Currency, Indicator, Measurement, and Funding.

Of course, the entities appropriate to a particular DMP would be selected and used.

These entities need to be linked by linking entities to provide the role relationship (semantics) between base entities and the temporal duration of the truth of the assertion (the role linking the base entities). The linking entities can refer to instances within the same base entity (e.g. Research Product related to Research Product: with role 'derived' or Research Product related to Organisation: with role 'rightsholder'). Concepts such as availability are a relationship between the Research Product and e.g. Organisation with an appropriate role (e.g. manager) and a temporal duration. A similar relationship exists between a Research Product and an Organisation in the form of a licence (role) with temporal duration.

This structure gives great flexibility: the role relationships between Research Product and Person could be creator, reviewer, user…; those between Research Product and Facility, Equipment and service record the digital collection of the asset (Research Product). Indicators and measurement relate to quality when linked to Research Product. The address information may be linked to an organisation (such as the one owning the facility), the facility itself, the person or the organization employing the person (for the purpose of research).

The metadata structure outlined above has been encoded – partially - in the CKAN metadata of EUDAT B2FIND/B2SAVE and – using RDF – could be made compatible with the W3C PROV-O[24] standard for provenance (so linking curation and provenance). Additionally, the above conceptual structure has been encoded in CERIF (Common European Research Information Format; an EU recommendation to the Member States) which is used widely for research information management but also for the

[24] https://www.w3.org/TR/prov-o/.

EPOS project where it forms the catalogue. The ongoing ENVRIplus rich metadata catalogue (CERIF) involves harvesting from EPOS and conversion of CKAN records from the ENVRIplus CKAN catalogue harvested from other RIs. CERIF has been mapped to DC (Dublin Core)[25], DCAT (Data Catalogue Vocabulary), CKAN (Comprehensive Knowledge Archive Network which has its own metadata format based on DC) and ISO19115/INSPIRE (an EU directive). The initial mapping to/from PROV-O has been done in joint work between euroCRIS and CSIRO, Canberra [9]. CERIF provides a 'switchboard' for interoperability as a superset model compared with the others, capable of representing a fully connected graph and having declared semantics with crosswalk capability [10, 11].

However, the existing metadata standards used within the RIs do not reach this level of richness of representation. Convertors have been provided from within the project and from other projects e.g. VRE4EIC[26], but RIs need to provide additional information, supplementing that in their existing metadata, to achieve appropriate curation (and for that matter, provenance and cataloguing) especially for interoperation purposes. For example, typical provenance information in metadata standards such as DC, DCAT, ISO19115 and others is human-readable text and not machine-understandable.

The chapter on cataloguing (Chapter 8) describes the catalogue implementation using CKAN and CERIF as the canonical metadata standard and implements them as a prototype.

### 5.2.3 Curation Processes

The processes associated with curation are:

1. Store an asset (e.g. dataset) with metadata sufficient for curation purposes;
2. Discover an asset using the metadata – the richer the metadata and the more elaborate the query the greater the precision in discovering the required asset(s);
3. Copy an asset with its updated metadata (to have a distributed backup version);
4. Copy an asset with its updated metadata (media migration to ensure availability)
5. Move an asset with its updated metadata (to a distributed location if the original location is unable to manage curation);
6. Partition an asset and copy/move across distributed locations with its updated metadata (for performance, privacy and security);
7. Partition an asset and copy/move across distributed locations with its updated metadata (for performance including locality of e.g. data with software and processing power)

The processes were defined based on the requirements solicited [6]. All these processes could be applied to a set of assets as well as a single asset. These processes are all simple given rich metadata in the catalogue as outlined above. The processes are documented and specified in the ENVRI RM (Reference Model).

[25] https://www.dublincore.org/.

[26] https://www.vre4eic.eu/.

## 6 Conclusion

The final design of the curation functionality aims to maximise flexibility while retaining compatibility with provenance and the catalogue. The catalogue is central to the design and implementation. The choice of the metadata elements in the catalogue (including their syntax and semantics) is crucial for the processes not only of curation but also of provenance and catalogue management and utilisation. The metadata model of the catalogue has also to permit interoperation among RIs as well as the usual processes associated with metadata catalogues: discovery, contextualisation and action. This implies that the model must be a superset (in the representation of syntax and semantics) of the metadata models used or planned within the RIs.

The chapter on cataloguing (Chapter 8) covers the implementation of CKAN (as used in EUDAT) and CERIF for the metadata catalogue.

This curation work relates closely to other tasks: cataloguing and provenance but also Identification and citation and processing leading towards representation in the reference model and the overall architecture design and evaluation.

The choice of a metadata standard for the catalogue was a critical decision for the project and the ability of RIs to compare CKAN and CERIF for cataloguing (related to the cataloguing processes of discovery, contextualisation and action), curation and provenance has been instructive.

The work on curation has caused the RIs to increase their attention to – and effort on – curation. RIs will now – with their DMPs – decide which assets to keep and curate, and which to delete and lose. The result of positive action is archives of curated environmental data essential for later research especially comparing the state of the environmental domain at that (future) time with now and past states as recorded. Some RIs need to store raw data to allow subsequent reprocessing/validation before interpretation. Reducing storage costs make this feasible but the cost of metadata generation is high and needs to be weighed against the benefits. Some RIs may be engaged in global collaborations, e.g. Euro-Argo or operate under global coordination, e.g. for atmospheric observations that need to be recognised by the IPCC[27]. The RIs need to fit their curation plans into this larger context and may even draw on the resources provided by that context. If these commitments to compatibility for curation demand only metadata and processes that are a subset of those proposed here, then interoperability and compatibility are assured. This will be clarified via DMPs, so that ENVRIplus can more accurately judge the residual requirement.

**Acknowledgements.** This work was supported by the European Union's Horizon 2020 research and innovation programme via the ENVRIplus project under grant agreement No. 654182.

## References

1. Zhao, Z., et al.: Knowledge-as-a-service: a community knowledge base for research infrastructures in environmental and earth sciences. In: 2019 IEEE World Congress on Services (SERVICES), pp. 127–132. IEEE, Milan (2019). https://doi.org/10.1109/SERVICES.2019.00041

[27] https://www.ipcc.ch/.

2. The Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System (OAIS), recommended practice CCSDS 650.0-M-2, June 2012 (2012). https://public.ccsds.org/Pubs/650x0m2.pdf. Accessed 01 Dec 2019
3. Using OAIS for Curation. DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3354. http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation. Accessed 01 Nov 2019
4. Atkinson, M., et al.: A consistent characterisation of existing and planned RIs. ENVRIplus Deliverable 5.1, submitted on 30 April 2016. http://www.envriplus.eu/wp-content/uploads/2016/06/A-consistent-characterisation-of-RIs.pdf. Accessed 01 Dec 2019
5. Myers, J., et al.: Towards sustainable curation and preservation: the SEAD project's data services approach. https://experts.illinois.edu/en/publications/towards-sustainable-curation-and-preservation-the-sead-projects-d. Accessed 01 Nov 2019
6. Jeffery, K., Asserson, A.: Supporting the research process with a CRIS. In: Asserson, A.G.S., Simons, E.J. (eds.) Enabling Interaction and Quality: Beyond the Hanseatic League; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, pp. 121–130 Leuven University Press (2006). ISBN 978 90 5867 536 1
7. Jeffery, K., et al.: Data curation in system level sciences: initial design. ENVRIplus deliverable report D8.1 (2017). http://www.envriplus.eu/wp-content/uploads/2015/08/D8.1-Data-Curation-in-System-Level-Sciences-Initial-Design.pdf
8. Dorbeva, M., Kim, Y., Ross, S.: Instalment on "Automated Metadata Generation". http://www.dcc.ac.uk/webfm_send/1513. Accessed 06 Jan 2020
9. Compton, M., Corsar, D., Taylor, K.: Sensor data provenance: SSNO and PROV-O together at last, In: Taylor, K., Gruetter, R. (eds.) Terra Cognita - Semantic Sensor Networks, TC-SSN 2014 - ISWC 2014. CEUR Workshop Proceedings, Trentino, Italy, pp. 67–82 (2014)
10. Martin, P., Remy, L., Theodoridou, M., Jeffery, K., Zhao, Z.: Mapping heterogeneous research infrastructure metadata into a unified catalogue for use in a generic virtual research environment. Future Gener. Comput. Syst. **101**, 1–13 (2019). https://doi.org/10.1016/j.future.2019.05.076
11. Remy, L., et al.: Building an integrated enhanced virtual research environment metadata catalogue. J. Electron. Libr. (2019). https://zenodo.org/record/3497056
12. Zhao, Z., et al.: Reference model guided system design and implementation for interoperable environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science, pp. 551–556. IEEE, Munich (2015). https://doi.org/10.1109/eScience.2015.41
13. Chen, Y., et al.: A common reference model for environmental science research infrastructures. In: Proceedings of EnviroInfo 2013 (2013)