D8.2
Data Curation Prototype

WORK PACKAGE 8 – DATA CURATION AND CATALOGUING

LEADING BENEFICIARY: NERC

Author(s):	Beneficiary/Institution
Keith G Jeffery	NERC
Zhiming Zhao	UvA

Accepted by: Zhiming Zhao (Theme 2 leader)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182

Deliverable type: [REPORT]

Dissemination level: PUBLIC

Deliverable due date: 31.10.2018/M42

Actual Date of Submission: 31.10.2018/M42.

ABSTRACT

Data curation is commonly under-resourced in research activity because it is not directly and visibly related to scientific production.

However, many important research discoveries have been made by re-working old data and/or by comparison of old data with recently collected data. This is particularly true of environmental sciences where understanding the atmospheric, biospheric, hydrospheric and geospheric processes usually require long-term observation and subsequent analysis. Also, many observations are unique in spatial and temporal coordinates and the opportunity to observe the same phenomenon will never recur.

Furthermore, validation and re-validation of research results requires open and understandable access to the data used in the preparation of the original publication.

Data curation is thus an important aspect of ENVRIplus and a key element of the ICT architectural and governance design. Data curation is integral to research methods (supporting, influencing, recording), workflows and processes and also integrates with all ICT activities through cataloguing and provenance. With an evolving policy of FAIR [Force11 2011] for open access to data – as well as publications – and, in time, software developed from the open source movement – curation has become more visible and necessary.

This deliverable describes the work of T8.1 initial and subsequent architectural design phases of ENVRIplus. And the development from the initial deliverable on curation (D8.1) to the current state.

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Malcolm Atkinson	University of Edinburgh
Robert Huber	Bremen University

Document history:

Date	**Version**
29.05.2018	Outline for comments
02.06.2018	Draft 1
31.07.2018	Draft 2 after WP8 review
23.10.2018	Draft 4 after 2 internal ENVRIplus reviews
24.10.2018	Proposed final version to Theme 2 leader

30.10.2018	Accepted by Zhiming Zhao

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors (Author names+email addresses)

TERMINOLOGY

A complete project glossary is provided online here:
https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonize policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance transdisciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

Contents

INTRODUCTION

Abstract

Data curation is commonly under-resourced in research activity because it is not directly and visibly related to scientific production.

However, many important research discoveries have been made by re-working old data and/or by comparison of old data with recently collected data. This is particularly true of environmental sciences where understanding the atmospheric, biospheric, hydrospheric and geospheric processes usually require long-term observation and subsequent analysis. Also, many observations are unique in spatial and temporal coordinates and the opportunity to observe the same phenomenon will never recur.

Furthermore, validation and re-validation of research results requires open and understandable access to the data used in the preparation of the original publication.

Data curation is thus an important aspect of ENVRIplus and a key element of the ICT architectural and governance design. Data curation is integral to research methods (supporting, influencing, recording), workflows and processes and also integrates with all ICT activities through cataloguing and provenance. With an evolving policy of FAIR [Force11 2011] for open access to data – as well as publications – and, in time, software developed from the open source movement – curation has become more visible and necessary.

This deliverable describes the work of T8.1 initial and subsequent architectural design phases of ENVRIplus and the development from the initial deliverable on curation (D8.1) to the current state.

Method

This activity (T8.1 within WP8) was undertaken by the primary author with contributions from key staff from other partners. The steps taken within the period M18-M42 of ENVRIplus are as follows:

1. Continuous review of D8.1 in the light of (a) evolving user requirements; (b) evolving technological perspectives external to the project; (c) evolving technological possibilities within the project
2. WP8 and wider discussion on the commonalities of metadata required and processes / workflows between curation and other ICT aspects particularly cataloguing and provenance but also identification and citation (WP6) and processing (WP7).
3. WP5-WP8 discussions on a representation of curation in the developing ENVRI Reference Model;
4. WP9-WP8 discussions on the evaluation of curation – particularly against the use cases;
5. Comparison of proposed curation architecture derived from D5.1 with that of the ENVRI RM;
6. Development of metadata and processing architecture for curation (together with cataloguing and provenance);
7. Design of governance for curation;

A key aspect is that no separate prototype software and governance system for curation has been developed. Instead, it was found to be beneficial to combine the development of curation solutions alongside those for cataloguing and provenance to ensure an integrated approach not least because they share many metadata entities/attributes/properties. This was recommended in the internal deliverable 'Implications of D5.1 for WPs 5,6,7,8'. Thus, the curation prototype is – in fact – a facet of the cataloguing system of ENVRIplus.

PROGRESS SINCE D8.1 (M18)

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).

As noted in D8.1, 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, with processing, with optimisation, with modelling and with architecture. This approach has been followed in the work leading to D8.2.

A key aspect of curation, noted in D8.1 and further supported during the work on curation leading to D8.2, 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 DMPs (Data Management Plans) 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.

Progress in 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 well as we are able by investing in curation has to be a key element of the ENVRI research culture with broad support by 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.

The curation stage of the lifecycle is also when metadata concerning quality is recorded. 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, the better judgement on

quality can be made. The quality processes for some ENV RIs have been studied in [Ma 2018] and both a quality taxonomy and potential improvements recommended.

There has been significant progress since D8.1 – a period of 24 months: (1) the RIs appreciate the curation lifecycle as described in D8.1; (2) the RIs 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.

Curation Lifecycle

The desirable lifecycle is represented by a DCC diagram [Figure 1] .

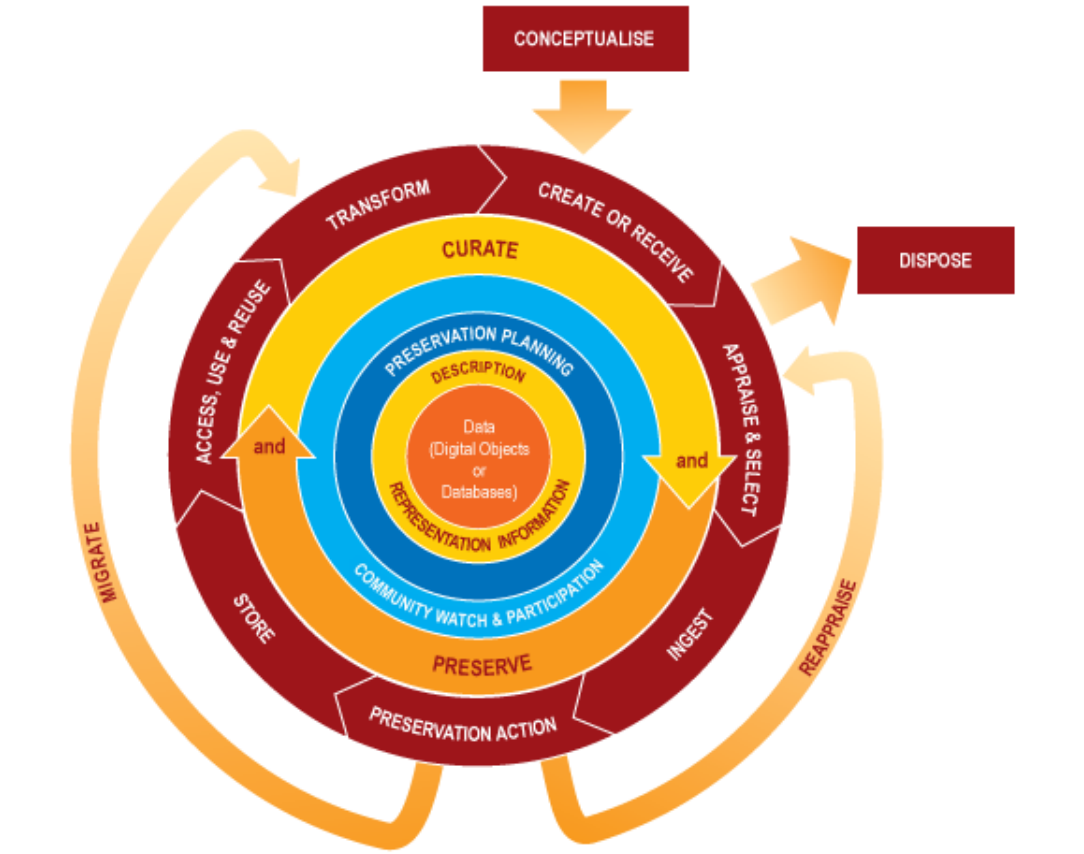

FIGURE 1: THE CURATION LIFECYCLE MODEL FROM DCC (THE DIGITAL CURATION CENTRE)

Data Management Plan

A DMP is defined (Wikipedia) "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".

The ENVRIplus RIs now have DMPs and utilise these as a basis for internal policy making, roadmapping, technological planning and governance of asset management, the latter within the framework of governance established by the RI e.g. the governance of an ERIC or a consortium through a consortium agreement.

OAIS Reference Model

As documented in D8.1, OAIS (Open Archival Information Systems Reference Model - ISO 14721:2002[1] - 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 the 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 [Higgins 2006]. Some RIs have considered OAIS as a framework but none has implemented it fully.

In order to populate such a framework a rich metadata element set is required. Much work M18-M42 has been done investigating various metadata standards to assess their suitability for curation (as well as for cataloguing and provenance). Within the work of RDA (Research Data Alliance) MIG (Metadata Interest Group) – of which the author is co-chair – a set of metadata elements in a structure for the purposes of curation, cataloguing and provenance according to FAIR[2] principles has been proposed[3].

RDA (Research Data Alliance)

The Research Data Alliance has groups working on this. Clearly there is benefit to ENVRIplus in alignment with the evolving RDA metadata recommendations which assist greatly not only in curation but also cataloguing, provenance 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[4]. RDA proposed some metadata principles which are now generally accepted:

- The only difference between metadata and data is 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 (relevance, quality, restrictions (rights, costs)) and for coupling users, software and computing resources to data (to provide a VRE);
- Metadata must be machine-understandable as well as human understandable for autonomicity (formalism);
- Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…);

[1] http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284
[2] https://www.force11.org/group/fairgroup/fairprinciples
[3] http://www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf
[4] https://www.rd-alliance.org/groups/metadata-ig.html

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, 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 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.

Problems to be Overcome

Some important problems – derived from D5.1 – were documented in D8.1. In that deliverable it was stated:

"These seven aspects of curation may be tackled incrementally, but ultimately 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 practicing domain scientists to help the ICT experts deliver relevant curation systems."

The table below documents the incremental progress achieved for each problem.

Problem to be overcome from D5.1 documented in D8.1	Work done M18-M42 in T8.1 and related WP8 and WP5 tasks
Motivation: 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.	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 in creased interest in and compliance with curation principles.
Business model: 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.	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: 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	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. The co-development of rich metadata cataloguing, curation and provenance in WP8 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.

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 [Myers *et al.* 2015]. That paper was presented at our ENVRIplus organized workshop at IEEE e-Science Conference, Munich in our IT4RIs workshop.	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 additionally to CKAN 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: 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 [JeAs 2006].	Within some RIs we see increasingly the use of workflows (and, indeed, in some, automation of workflow deployment across multicloud 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[5] where the author was scientific coordinator.
Curation of data: 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 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. Derivatives may be generated and require management including relationships with the	Over the last 24 months 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

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

original dataset and all its attendant metadata.	efficiency overcoming latency – in a multicloud environment is being developed in the MELODIC project[6] where the author is a consultant to the project.
Curation of software: 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 becomes 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.	The issue was novel to most RIs when introduced in T8.1 and recorded in D8.1. The requirement is now appreciated but the metadata systems in use in most RIs are incapable of providing a technological solution. It is further complicated because many developers – including those in some RIs – use GitHub 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 catalog the software and its evolution and monitor work elsewhere that will provide appropriate solutions.
Curation of operational environments: 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 accuracy, precision, operational speed, capacity and many more. The operating system has defined characteristics and includes device drivers – i.e., a software library	The issue was novel to most RIs when introduced in T8.1 and recorded in D8.1. 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

[6] http://melodic.cloud/

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.	external developments to find solutions as they are developed. Increasingly, there appears to be a partial solution in containerisation using e.g. Docker[7] or Kubernetes[8]. Unlike 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).
Problem to be overcome arising M18-42	
Curation of 'raw' data collected by sensors or instruments	While the requirements from D5.1 concentrated on curation of validated or part-processed data, some RIs require to curate (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 has distributed observatories with differing policies. In contrast EuroARGO 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 the 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 or SSNO) and e.g. linked by PIDs with the incoming data stream[9].

[7] https://www.docker.com/
[8] https://kubernetes.io/
[9] https://www.rd-alliance.org/group/persistent-identification-instruments/case-statement/persistent-identification-instruments

A longer-term horizon

In D8.1 it as claimed that there is some cause for optimism. This list of reasons is supplemented with comments relating to the work M18-M42:

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 n 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 WP8 within ENVRIplus but also international efforts such as RDA and the FAIR initiative.
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.

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. In D8.1 it was stated that they may reach production status in this timeframe[10]. At present – although progress has been made – there are no appropriate systems although research indicates some cause for optimism.

Issues and implications

D8.1 listed issues and implications. Progress M18-M42 is recorded alongside them in the following table.

Issues and Implications from D8.1	**Work M18-M42**
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;	The joint work especially with T8.2 cataloguing - and following the recommendations of D8.3 and D5.5 – has led to the development of two catalogues, one using CKAN as in EUDAT B2FIND and the other using CERIF as used in EPOS. Experiments are underway to evaluate the two approaches for capability as the core metadata scheme.

[10] http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/automated-metadata-extraction

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;	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 from digital objects shows promise but production system readiness is some years away. However, metadata provision from equipment-generated streamed data is available;	This has been monitored but the current systems are not yet at production status sufficient to be recommended to the RIs.
ENVRIplus should adopt the DCC recommendations;	Following acceptance of D8.1 this is achieved. However, implementation is incremental.
ENVRIplus should track the relevant RDA groups and – ideally – participate	Following acceptance of D8.1 both tracking and participation are pursued actively. Of particular relevance is the work on the RDA Metadata Element set which could be a candidate for a common metadata scheme.
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.	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.

USE CASES AND REQUIREMENTS

Introduction

Use cases were used in the production of D8.1. Since that time curation requirements have changed little (although requirements are being tracked through site visits and ENVRI meetings) but provision and adoption of governance and technical solutions to those requirements have been disappointingly slow.

Requirements from Use Cases

The *Curation* requirements from seven RIs were documented in the preparation work of D5.1; see the wiki page for details[11]. It was clear that the requirements were already conceptually and practically interrelated with *Cataloguing* and *Provenance* in WP8. As remarked above, it should also strongly couple with the work on *Data Identification and Citation (WP6)*. Further issues arose in the analysis of requirements and were documented in D5.1. Here these are tabulated and the work to address them M18-M42 is described for each below.

Further Issues to be Addressed

Issue Identified in D5.1	**Work M18-M42 to address the issue**
The appreciation of the needs for *Curation* is varied and often limited, one manifestation of this is the almost universal absence of complete data management plans [12]. In practice a DMP evolves providing early the essentials for data collection and availability to the immediate community and later interoperability across the whole domain with enhanced metadata including not only descriptions of the data but also information on rights, security and privacy. Consequently, this topic again poses a requirement for an ENVRIplus programme of awareness raising and training. If that is conducted collaboratively then it will also help develop cross-disciplinary alliances that will benefit scientific outcomes, management decisions and long-term cost-benefit trade-offs.	The appreciation of curation has increased significantly, assisted by scientific use cases where the re-use of old assets has proved to be not only beneficial but essential. The understanding and appreciation of the data curation cycle (within the research lifecycle) has improved considerably, not least because of repeated presentations on the subject in project meetings and workshops. There has been an increase in RIs generating DMPs to assist in clarifying requirements and defining both technological and governance solutions. Increasingly – not only for curation but also for cataloguing and provenance – RIs understand the need for rich metadata due to presentations (awareness and training) at ENVRI events.
The need for intellectual as well as ICT interworking between these closely related topics: *Identification and Citation*, *Curation*, *Cataloguing* and *Provenance* is already recognised. Their integration will need to be well supported by tools, services and	Within WP8 there has been close interworking between T8.1 (curation), T8.2 (cataloguing), T8.3 (provenance) and also with Identification and Citation depending on Unique Identifiers.

[11] https://wiki.envri.eu/display/EC/Curation+requirements

[12] These may be latent in policy and management documents of each RI. Drawing them together into a formal DMP will take time. It might benefit from being collaborative, and from training such as that offered by the DCC, http://www.dcc.ac.uk/.

processing workflows, used to accomplish the scientific methods and the *Curation* procedures. However, there was negligible awareness of the need to preserve software and the contextual information necessary to re-run it with identical effects - or with well-understood, controlled and intended variations. The need for this combination for reproducibility is identified by Belhajjame *et al*. with implementations automatically capturing the context and synthesising virtual environments [Belhajjame 2015].	The key is rich metadata and work in T8.2 on both CKAN (as used in EUDAT B2FIND) and CERIF (as used in EPOS for example) is exposing the need for rich metadata for these functions. There is now awareness of the need to preserve software (or better, software specifications) and operational environments including software libraries – especially for scientific reproducibility. Again this highlights the need for rich metadata ensuring that the assets are re-usable (by humans and by autonomic IT systems) for as long as required.
As above, it is vital to support the day-to-day working practices and the innovation steps that occur in the context of *Curation* with appropriate automation and tools. This is critical both to make good use of the time and effort of those performing *Curation*, and to support innovators introducing new scientific methods with consequential *Curation* needs.	The tool support is bound intimately with the support for (a) cataloguing: used for discovery, contextualisation (relevance and quality) and (re-)use; (b) provenance: used for audit, reproducibility and contextualisation. Each RI has its own technological and governance system(s). The Theme 2 WPs have developed recommendations and prototype solutions but adoption depends on the RIs.
The challenge of handling all forms of data described in 'Problems to be overcome' for *Identification and Citation* is compounded with the need to properly capture diverse forms of software (or, better, formal specifications of the software) and a wide variety of, often distributed, computational contexts in order to fully support reproducibility.	This again requires rich metadata and hence the work on this topic M18-M42 has been directed to this end. Investigation and documentation of the various local (and not generally used more widely) extensions to standards such as DC (Dublin Core), CKAN (Comprehensive Knowledge Access Network) and ISO19115/INSPIRE all indicate this need.
Curation needs to address preservation and sustainability; carefully preserving key information to underwrite the quality and reproducibility of science requires that the information remains accessible for a sufficient time. This is not just the technical challenge of ensuring that the bits remain stored, interpretable and accessible. It is also the socio-political challenge of ensuring longevity of the information as communities' and	This aspect has been emphasised in presentations within ENVRIplus and by ENVRIPlus to other audiences. There may well be a scientific case for curation in perpetuity - both for reproducibility and for longitudinal time-based studies - but the economic implications of a curation policy and technology with an infinite timespan are unacceptable. There are also questions over the cost of curation compared with re-

funders' priorities vary. This is a significant step beyond archiving, which is addressed in EUDAT with the B2SAFE service[13].	collection of data (observation, experiment or simulation). Of course, for many environmental science situations re-collection is impossible or difficult.
One aspect of the approach to sustainable archiving is to form federations with others undertaking data curation, as suggested by OAIS[14]. Federation arrangements are also usually necessary in order that the many curated sources of data environmental scientists need to use are made conveniently accessible. Such *data-intensive federations* (DIF) underpin many forms of multi-disciplinary collaboration and supporting them well is a key step in achieving success. As each independently run data source may have its own priorities and usage policies, often imposed and modified by its funders, it is essential to set up and sustain an appropriate DIF for each community of users. Many of the RIs deliver such federations, *today without a common framework to help them*, and many of the ENVRIplus partners are members of multiple federations.	The DIF approach has been discussed at ENVRI meetings and appears in other deliverables. The advantages of having multiple copies of assets is clear both for availability (preservation) and local availability (performance). However, DIFs increase the requirement for rich metadata and consistency among metadata catalogues to ensure appropriate access paths and provenance information.

[13] http://www.eudat.eu/b2safe

[14] http://wiki.dpconline.org/index.php?title=6-3

ARCHITECTURAL DESIGN PRINCIPLES FOR CURATION

Introduction

D8.1 asserted three aspects of the then current state. Each, below, is supplemented by the work done in M18-M42:

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 are 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).

Analysis

Further work between D8.1 and D8.2 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[15] (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.

Increasingly we live in a world where data or information may be altered to fit new political 'facts'. 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'.

[15] https://eugdpr.org/

Recommendation

It is clear that in the period M18-M42 the RIs have appreciated the need for a common rich metadata standard covering not only curation but also cataloguing and provenance. The requirement for protection of personal data and assurance of fixity underlines the need for rich metadata appropriate for enforcing access control. The WP8 team has been working towards this and has been evaluating the solutions described in D8.1 and D8.2 within the context of the D5.1 requirements and D5.5 architecture.

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

GOVERNANCE PRINCIPLES FOR CURATION

Introduction

In the period M18-M42 there has been a significant increase in governance activity related to curation. Notably this has involved the use of DMPs.

Recommendation

The key recommendation from D8.1 is that RIs should have a DMP, and ideally use the DCC documentation. This is now achieved across many RIs and DMPs are being used to manage the governance of curation and the management of (at least some of) the technological aspects of curation. Richer metadata and appropriate procedures are required, based on the governance of the RI.

RELATIONSHIP TO THE ENVRI-RM

Introduction

The ENVRI-RM provides a formal method for describing the common information structures and operations of the RIs within ENVRI both existing and necessary to reach the objectives of ENVRIplus. In the case of curation, the key information is in the Information viewpoint[16] and suggests the steps: data acquisition, backup, assign identifier, add metadata, annotate data, annotate metadata, build conceptual model, global conceptual models before moving on into data publishing.

Analysis

At the time of 8.1 the ENVRI-RM was partly developed and was insufficiently developed for curation purposes. Intensive work with colleagues responsible for D5.2 led to a much better representation of curation in the RM.

Next Steps

Further refinements of the RM continue but the major work has been done M18-M42.

[16] https://wiki.envri.eu/display/EC/IV+Lifecycle+in+Detail#IVLifecycleinDetail-DataCuration

FINAL DESIGN

Introduction

The initial design 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 catalog metadata and the curation processes. The final design confirms the initial design and adds detail.

Catalog Metadata

The catalog – 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 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 T8.2 and its deliverable, D8.4.

However, the required metadata elements can be specified, derived from D5.1 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 organization (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[17] standard for provenance (so linking curation and provenance). Additionally, the above conceptual structure has been encoded in CERIF (Common European Research Information Format; a EU recommendation to Member States)[18] which is used widely for research information management but also for the EPOS project where it forms the catalog. The ongoing ENVRIplus rich metadata catalog (CERIF) involves harvesting from EPOS and conversion of CKAN records from the ENVRIplus CKAN catalog harvested from other RIs. CERIF has been mapped to DC (Dublin Core), DCAT (Data Catalog Vocabulary), CKAN (Comprehensive Knowledge Archive Network which has its own metadata format based on DC) and ISO19115/INSPIRE (a EU directive). The initial mapping to/from PROV-O has been done in joint work between euroCRIS and CSIRO, Canberra. 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.

However, the existing metadata standards used within the RIs do not reach this level of richness of representation. Convertors have been provided, 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.

D8.4 from T8.2 describes the catalog implementation using CKAN and CERIF as the canonical metadata standard and implements them as a prototype.

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)

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 catalog as outlined above. The processes are documented and specified in the ENVRI RM.

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

[18] http://www.eurocris.org/cerif/main-features-cerif

CONCLUSIONS

The final design of the curation functionality aims to maximize flexibility while retaining compatibility with the other tasks in WP8, namely provenance and the catalog. The catalog is central to the design and implementation. The choice of the metadata elements in the catalog (including their syntax and semantics) is crucial for the processes not only of curation but also of provenance and catalog management and utilisation. The metadata model of the catalog has also to permit interoperation among RIs as well as the usual processes associated with metadata catalogs: discovery, contextualisation and action. This implies that the model must be a superset (in representation of syntax and semantics) of the metadata models used or planned within the RIs.

D8.4 describes the T8.2 implementation of CKAN (as used in EUDAT) and CERIF for the metadata catalog.

IMPACT ON THE PROJECT

This deliverable relates closely to other tasks and deliverables, first within WP8 (cataloguing and provenance) but also WP6 (Identification and citation) and WP7 (processing) leading towards representation in the reference model and the overall architecture design (WP5) and evaluation (WP9).

The choice of metadata standard for the catalog 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 T8.1 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., EuroARGO or operate under global coordination, e.g., for atmospheric observations that need to be recognized by the IPCC. The RIs need to fit their curation plans into this larger context and may even draw on 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.

IMPACT ON STAKEHOLDERS

The major impact on stakeholders is archives of well-curated assets for subsequent (re-)use. The correct choice of catalog metadata standard has a huge influence on stakeholders since it conditions what processing facilities are available to all RIs in ENVRIplus. The metadata has to support not only curation and provenance but also the usual research processes of discovery, contextualization (which may involve visualisation) and action which utilizes the catalog to access and use the digital assets of the RIs and – more importantly perhaps – to interoperate across the RIs to allow novel interdisciplinary research.

This deliverable should cause RIs to continue with their strategy for curation (started with D8.1) and increase attention and effort on it, not only for the benefit of their community now and in the future but also for other communities interoperating with their own to achieve cross-domain research results. For some RIs, developing their DMP further may stimulate this process and provide opportunities for collaboration and education.

REFERENCES

[Belhajjame 2015] K. Belhajjame, J. Zhao, D. Garijo, K. Hettne, R. Palma, O. Corcho, J.-M. Gómez-Pérez, S. Bechhofer, G. Klyne, and C. Goble, "*A Suite of Ontologies for Preserving Workflow-Centric Research Objects*," Journal of Web Semantics, 2015.

[Force11 2011] https://www.force11.org/group/fairgroup/fairprinciples

[Higgins 2006] "Using OAIS for Curation". DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3354. Available online: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

[JeAs 2006]Keith G Jeffery, Anne Asserson: '*Supporting the Research Process with a CRIS*' in Anne Gams Steine Asserson, Eduard J Simons (Eds) '*Enabling Interaction and Quality: Beyond the Hanseatic League*'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp 121-130 Leuven University Press ISBN 978 90 5867 536 1

[MA2018] Maduro, Jordan, 2018, Towards a taxonomy for quality control in environmental sciences, master thesis, University of Amsterdam, Zenodo, doi:10.5281/zenodo.1419494.

[Myers et al 2015] James Myers, Margaret Hedstrom, Dharma Akmon, Sandy Payette, Beth A. Plale, Inna Kouper, Scott McCaulay, Robert McDonald, Isuru Suriarachchi, Aravindh Varadharaju, Praveen Kumar, Mostafa Elag, Jong Lee, Rob Kooper, Luigi Marini: 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

Appendices

Appendix 1: Proposed Questions to ascertain the state of curation in any RI

1. is it possible to recover/read/act upon a dataset with a given name or keywords and version and date of curation?
2. is it possible to recover/read/act upon a software module with a given name or keywords and version and date of curation?
3. is it possible to recover/read/act upon a workflow with a given name or keywords and version and date of curation?
4. for all the above ideally with rights (e.g. licence) and associated
5. organisations or persons (e.g. rights holder)
6. for all the above is it possible to see the positioning and relationships of the object within a network of information such as previous and subsequent versions, related datasets or software to a given dataset, related organisation or person to a given object.....(this is where curation meets provenance).
7. Is the location of the dataset or other digital object known and its locality with respect to other relevant datasets and other digital objects (e.g. software, computing resources) so that workflow may be optimised

And finally:

8. is a current and acceptable (sustainable) DMP (data management plan) in place