D8.1
Data Curation in System Level Sciences: Initial Design

WORK PACKAGE 8 – DATA CURATION AND CATALOGUING

LEADING BENEFICIARY: NERC

Author(s):	Beneficiary/Institution
Keith G Jeffery	NERC
Zhiming Zhao	UvA
Barbara Magagna	EAA
Abraham Nieva de la Hidalga	Cardiff University
Leonardo Candela	CNR
Carl-Frederik Enell	EISCAT
Margareta Hellstrom	Lund University
Alex Hardisty	Cardiff University
Charles Paxton	University of S Andrews
Frank Toussaint	DKRZ

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

Accepted by: Zhiming Zhao (Theme 2 leader)

Deliverable type: [REPORT]

Dissemination level: PUBLIC

Deliverable due date: 31.10.2016/M18

Actual Date of Submission: 31.01.2017/M21.

ABSTRACT

Data curation is commonly the 'Cinderella' of ICT (Information and Communication Technologies). Usually, it receives little attention from researchers or managers and may be seen as a tedious chore to be done in wrapping up the research activity. Since research may well be continuous, such wrapping up may not occur.

In contrast, 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 requires long-term observation and subsequent analysis.

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 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 reviews the state of the art and recommends architectural principles to be taken into account (along with the inputs on other topics) in the initial and subsequent architectural design phases of ENVRIplus.

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Malcolm Atkinson	University of Edinburgh
Alex Vermuelen	Lund University

Document history:

Date	**Version**
20.05.2016	Outline for comments
12.10.2016	Corrected version to Theme 2 for comments
11.11.2016	Corrected version to internal review
15.11.2016	Modified version following first internal review
06.12.2016	Modified version following comments from WP8 and at

	ENVRIweek
02.01.2017	Version aligned with changes in D8.3 and D5.4
10.01.2017	To second internal review
14.01.2017	Modified version following internal review 2 and comments from the RM team
22.01.2017	Final modifications from suggestions received from Zhiming Zhao
22.01.2017	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.

TABLE OF CONTENTS

Contents

ABSTRACT 3
DOCUMENT AMENDMENT PROCEDURE 4
TERMINOLOGY 4
PROJECT SUMMARY 4
TABLE OF CONTENTS 5
INTRODUCTION 8
Abstract 8
Method 8
STATE OF THE ART AND REVIEW 10
Introduction, context and scope 10
Sources of state of the art technology information used 10
Short term analysis of state of the art and trends 10
A longer-term horizon 15
Issues and implications 15
USE CASES AND REQUIREMENTS 16
Introduction 16
Requirements from Use Cases 16
Further Issues to be Addressed 16
ARCHITECTURAL DESIGN PRINCIPLES FOR CURATION 18
Introduction 18
Recommendation 18
GOVERNANCE PRINCIPLES FOR CURATION 19
Introduction 19
Recommendation 19
RELATIONSHIP TO THE ENVRI-RM 20
Introduction 20
Analysis 20
Next Steps 21
INITIAL DESIGN 22
Introduction 22
Catalog Metadata 22
Curation Processes 23
CONCLUSIONS 24

IMPACT ON THE PROJECT 25
IMPACT ON STAKEHOLDERS 26
REFERENCES 27
Appendices 28

INTRODUCTION

Abstract

Data curation is commonly the 'Cinderella' of ICT (Information and Communication Technologies). Usually, it receives little attention from researchers or managers and may be seen as a tedious chore to be done in wrapping up the research activity. Since research may well be continuous, such wrapping up may not occur.

In contrast, many important research discoveries have been made by re-working old data and/or by comparison of old date with recently collected data. This is particularly true of environmental sciences where understanding the atmospheric, biospheric, hydrospheric and geospheric processes usually requires long-term observation and subsequent analysis. Clearly the quality of the data is important, and the curation process includes quality control. However, the quality required (and the justification of the associated cost) depends less on the actual curation and more on the re-use purpose. For some purposes very high quality data (and metadata) is required, for others acceptable research can be done with lower quality.

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 (support, influencing, reporting), workflows and processes and also integrates with all ICT activities through cataloguing and provenance. With an evolving policy of 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 reviews the state of the art and recommends architectural principles to be taken into account (along with the inputs on other topics) in the initial and subsequent architectural design phases of ENVRIplus.

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 first 18 months of ENVRIplus are as follows:

1. Literature review on curation and review of activities in other recent and current projects;
2. Extraction of curation requirements from use cases and stated requirements particularly from the work associated with D5.1 (note in parallel an exercise to extract requirements and update the ENVRI-RM was undertaken leading to D5.2);
3. 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).
4. WP5-WP8 discussions on representation of curation in the developing ENVRI Reference Model;
5. WP9-WP8 discussions on evaluation of curation – particularly against the use cases;

6. Comparison of proposed curation architecture derived from D5.1 with that of the ENVRI RM;
7. Initial design of metadata and processing architecture for curation
8. Initial design of governance for curation

STATE OF THE ART AND REVIEW

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).

It should be noted that 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.

The origins of curation stretch back to the earliest librarianship (including making copies to be distributed in monasteries and in the well-known case of the 'Magna Carta' in the UK with distribution to cathedrals – this finds its modern equivalent in LOCKSS[1]) and to the identification and cataloguing with metadata of objects of interest in museums.

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. 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.

Sources of state of the art technology information used

Relevant major sources are the Data Curation Centre (DCC), Open Archival Information System (OAIS) (both discussed below) and Research Data Alliance (RDA), which has several relevant groups notably preservation[2] but also active data management plans[3] and reproducibility[4]. Knowledge of the bibliography and of curation activities in current and recent projects provide further source material. The ENVRI RM (Reference Model) defines curation operations.

Short term analysis of state of the art and trends

The ideal curation culture will ensure – via an appropriate system - 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 current practice commonly falls far short of this with preservation commonly linked with backup or recovery (usually limited to the physical preservation of the digital asset) and lacking the steps of curation (selection, ingestion,

[1] http://www.lockss.org
[2] https://rd-alliance.org/groups/preservation-e-infrastructure-ig.html
[3] https://rd-alliance.org/groups/active-data-management-plans.html
[4] https://rd-alliance.org/groups/reproducibility-ig.html

preservation, archiving [including metadata]) and maintenance. Furthermore, in the current state while datasets may be curated it is rare for software or operational environments to be curated. Including these is necessary to achieve reusability [Belhajjame 2015]. Collecting them automatically has been demonstrated by [Santana-Perez 2016], where processes in a virtual environment are monitored and their interactions with external resources recorded. The collected information is used to automatically create a virtual image in which the job can be deployed and re-run on the cloud. However, while this is feasible in a homogeneous environment it is a leading-edge research topic to achieve this in a heterogeneous environment such as 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.

.

Curation Lifecycle

The desirable lifecycle is represented by a DCC (Digital Curation Centre) diagram [Figure 1]).

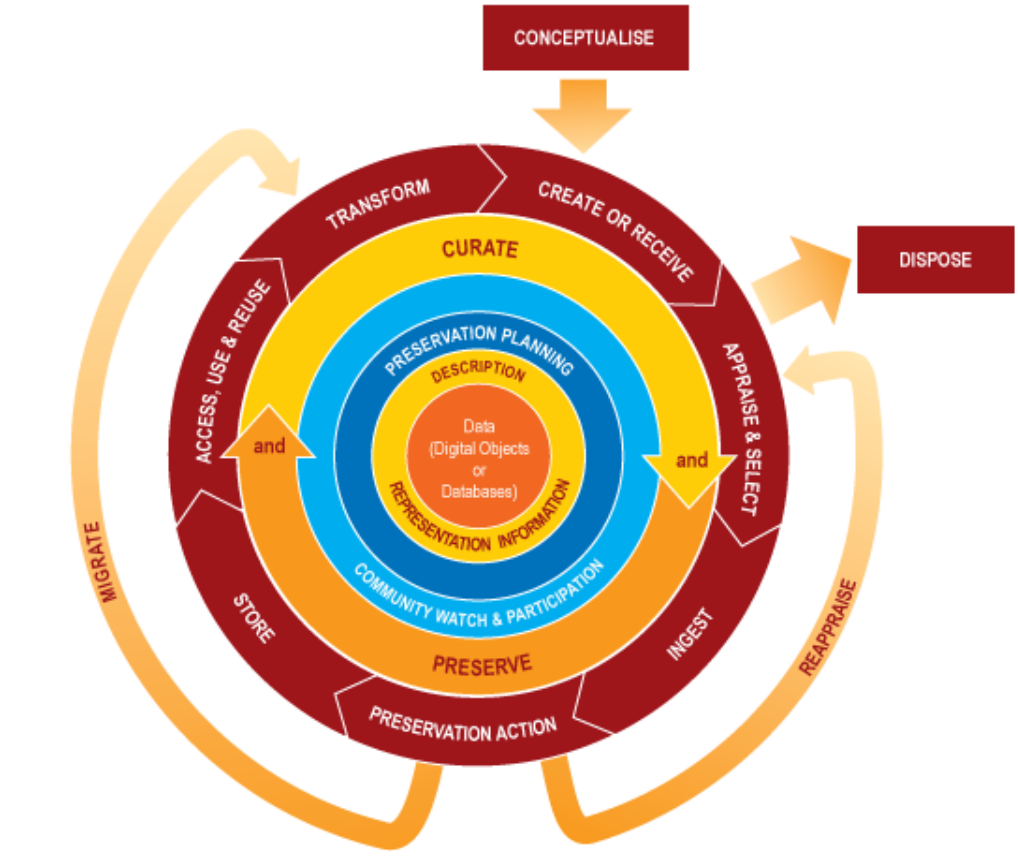

FIGURE 1: THE CURATION LIFECYCLE MODEL FROM DCC (THE DIGITAL CURATION CENTRE)

Data Management Plan

Increasingly research funders are demanding a DMP (Data Management Plan). Different organisations have proposed different templates and tools for plans but that of DCC is used widely[5] as is the US equivalent[6]. 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".

OAIS Reference Model

OAIS (Open Archival Information Systems Reference Model - ISO 14721:2002[7] - 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]. A set of metadata elements in a structure has been proposed[8].

RDA (Research Data Alliance)

The Research Data Alliance has groups working on this (see above). However, their work is brought together with that of other groups in the specification of metadata[9]. RDA has proposed some metadata principles:

- 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…)

And furthermore, a metadata element set that covers all the uses of metadata (not just curation):

- Unique Identifier (for later use including citation)

[5] http://dmponline.dcc.ac.uk
[6] http://dmp.cdlib.org
[7] http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284
[8] http://www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf
[9] https://www.rd-alliance.org/groups/metadata-ig.html

- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Spatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- 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) and so are not simple attributes with values. The RDA groups are currently working on 'unpacking' the elements to a form suitable for discovery, contextualisation and action by both humans and computers.

Problems to be Overcome

The following are some important problems – derived from D5.1 - that need to be addressed for curation:

1. **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.
2. **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.
3. **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 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.

4. **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 [Jeffery 2006].
5. **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 original dataset and all its attendant metadata.
6. **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.
7. **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 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.

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.

A longer-term horizon

There is some cause for optimism:

1. Media costs are decreasing – so more can be preserved for less (and the cost reduction hopefully matches the expansion of volume);
2. Awareness of the need for curation is increasing; partly through policies of funding organisations and partly through increased responsibility of some researchers;
3. Research projects in ICT are starting to produce autonomic systems that could be used to assist with curation.

However, the major problem is the cost of collecting metadata for curation. Firstly, incremental collection along the workflow with re-use of existing information should assist. Workflow systems should be evolved to accomplish this. Secondly, improving techniques of automated metadata extraction from digital objects may reach production status in this timeframe[10].

The complexity and changes in the natural world combined with human ingenuity developing sciences and their methods makes it difficult to predict which curated information will be needed and how it will be used. Information latent in observations or derivatives may become important. Methods may be re-used with significant updates to some of their components. The complexity of attempting to handle an open-ended set of possibilities has to be mitigated by agreed simplifications. Judging such tradeoffs is a key aspect of a curator's art. Developing DMPs in consultation with domain leaders, innovative scientists and funding stakeholders brings that judgement into play and records the developing agreements.

Issues and implications

1. 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;
2. 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;
3. 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;
4. ENVRIplus should adopt the DCC recommendations;
5. ENVRIplus should track the relevant RDA groups and – ideally – participate.
6. 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.

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

USE CASES AND REQUIREMENTS

Introduction

All the requirements obtained from the interviews and the use cases indicated some awareness of the need for digital curation. However, few RIs had advanced towards providing systems to achieve curation and even those that had advanced had not a full data management plan (including business case) in place.

Requirements from Use Cases

The *Curation* requirements validate the need for ENVRIplus developing curation solutions but do not converge on particular requirements. This brief analysis is based on the information supplied by seven RIs who responded to this topic during the preparation of D5.1; see the wiki page for details[11]. In the planned work of ENVRIplus this work is 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)*. Consequently, many of the issues that emerge are similar to those identified above. However, some further issues arise.

Further Issues to be Addressed

These are enumerated below:

1. 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.
2. 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 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].
3. 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.

[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/.

4. 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.
5. 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 funders' priorities vary. This is a significant step beyond archiving, which is addressed in EUDAT with the B2SAFE service[13].
6. 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.

[13] http://www.eudat.eu/b2safe
[14] http://wiki.dpconline.org/index.php?title=6-3_

ARCHITECTURAL DESIGN PRINCIPLES FOR CURATION

Introduction

We start with the state of the art and the requirements (see above). These indicate:

1. Technologies are available for curation but they may not be compatible with those for cataloguing and provenance;
2. Governance principles for curation are lacking widely among the ENVRI community;
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;

On the other hand, there are examples of successful curation activity in other communities such as those that support life-sciences reference data, e.g., PDB [Berman 2008] and those that support sky surveys [Szalay 2008].

Recommendation

A major problem in ENVRIplus is the heterogeneity of the RIs in both governance and technology. This may hinder take-up of any recommended curation solution. On the other hand, some of the RIs already form informal clusters (usually by environmental domain) so there may be scope for collaborative work on curation with consequent increased benefit for the cost.

The architecture described will steer the implementation of curation in ENVRIplus Theme 2. This will be an exemplar to illustrate the potential benefits. It may be adopted by clusters of RIs and developed to meet their needs or may reveal operational evidence that suggests future refinements. As curation requires system longevity a development and operational plan will need resources to achieve sustainability.

GOVERNANCE PRINCIPLES FOR CURATION

Introduction

Since there is (relative to the volumes of data) little curation activity in ENVRIplus RIs at present (as recorded in D5.1) we can make recommendations expecting that there will be few problems integrating with existing governance and practices.

Recommendation

The key recommendation is that RIs should have a DMP, and ideally use the DCC documentation. This ensures:

1. The RIs are actually thinking about the issue;
2. They generate governance recommendations relevant to their community and the assets of data, software and processing environments;
3. They adopt an appropriate metadata standard for cataloguing, curation and provenance or – in the event of multiple established heterogeneous metadata formats – they choose a canonical format and provide convertors;
4. Where appropriate RIs within an environmental domain cluster so as to share the cost (and benefit from experience) of curation;
5. The RIs consider quality control - and the cost - relative to the predicted benefits from the asset and its metadata for (re-)use.

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[15] 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

The ENVRI-RM does not cover an important aspect emerging from the work on D8.1 namely, the importance of the DMP to define processes and governance and to encourage awareness of the curation need. However, D5.2 indicates that it will be taken into consideration in future. The implication of the ENVRI-RM process steps is that curation follows immediately acquisition whereas it may well be done at some considerable time later than acquisition and possibly by a skilled curator rather than the data owner or data loader. In fact, much data may be acquired from equipment without humans being able to interfere with curation metadata during acquisition: this is beneficial for recording metadata associated with the measurements in the dataset but less helpful for wider utilisation (particularly many years later) of the data when more appropriate metadata for curation is required.

The distinction between metadata and annotation of data or metadata) may not be useful; in both cases annotation is defined as enrichment and – depending on the acquisition/curation process of a particular RI with a particular DMP – the enrichment may be continuous or sporadic. Indeed, it is likely the enrichment relates to provenance information (and therefore we need equivalencing of metadata elements across curation and provenance) and with the steps of the process and the resulting assets being recorded in the catalog thus confirming the intersection of metadata elements among these three aspects of WP8. D5.2 section 5.3.2 defines annotation as linking the metadata to conceptual models (within the context of semantic harmonisation). Semantic harmonisation requires a (conceptual) canonical rich metadata catalog; whilst this is recognised as a requirements in D5.2 it is planned future work to be represented in the ENVRI-RM.

The ENVRI-RM appears not to cover the curation (nor provenance and cataloguing) of software modules, workflows or computing environments all of which are necessary for curation. The relationships between these components are very important especially for reproducibility of research results and – if required – 'data rescue' where data with little or no metadata is rescued (including probably media migration).

Media migration as a process is also not mentioned in the ENVRI-RM yet in large datacentres a considerable proportion of time on automated robots managing storage is devoted to media migration to ensure availability of the digital objects recorded.

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

Next Steps

Having derived the lists of requirements, characterisations of existing state, characterisation of plans and issues from the work associated with D5.1 we can now move towards formalising this in the ENVRI-RM, building upon the basic structure for curation in the RM referenced above. D5.2 is a first step. The metadata and processes described in the initial design (below) form the basis for the IV (Information Viewpoint) and CV (Computational Viewpoint) respectively and can be brought together – after further consideration – in the EV (Engineering Viewpoint).

INITIAL DESIGN

Introduction

The initial design is 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.

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[16]. 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, D8.3.

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[17] (see above under 'State of the Art'). 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, 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').

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).

[16] https://www.force11.org/group/fairgroup/fairprinciples
[17] https://www.rd-alliance.org/groups/metadata-ig.html

The metadata structure outlined above could be encoded in RDF (as in the CKAN metadata of EUDAT B2FIND/B2SAVE) and – using RDF – could be made compatible with the W3C PROV-O[18] standard for provenance (so linking curation and provenance). Alternatively, the above conceptual structure could be encoded in CERIF (Common European Research Information Format; a EU recommendation to Member States)[19] which is used widely for research information management but also for the EPOS project where it forms the catalog. 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 can be provided, but it is certain that RIs will 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.

D8.3 from T8.2 proposes that ENVRIplus recommends CKAN and CERIF as the canonical metadata standard and implements them within any 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 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.

[18] https://www.w3.org/TR/prov-o/
[19] http://www.eurocris.org/cerif/main-features-cerif

CONCLUSIONS

The initial 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 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.

D8.3 from T8.2 proposes the use of CKAN (as used in EUDAT) and CERIF for the metadata catalog. However, before proceeding to detailed design and prototype implementation it is necessary to validate both CKAN and CERIF against the requirements of provenance (which task – T8.3 - does not start until later in the project) and the catalog – T8.2 where he work is parallel to curation - T8.1.

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 is a critical decision for the project.

It is expected that this deliverable will cause RIs to increase their attention to – and effort on – curation. RIs will – with their DMPs – decide which assets to keep and curate, and which to delete and lose. The result of positive action will be 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 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. These may need to fit their curation plans into this larger context and may even draw on resources it provides. 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 re-assess their strategy for curation 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 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.

[Berman 2008] Berman, H. M. *"The Protein Data Bank: a historical perspective" (PDB). Acta Crystallographica Section A: Foundations of Crystallography* **A64** (1): 88–95 (Jan. 2008). doi:10.1107/S0108767307035623

[Jeffery 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

[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

[Santana-Perez 2016] Idafen Santana Perez PhD thesis "*Conservation of Computational Scientific Execution Environments for Workflow-based Experiments Using Ontologies",* January 2016, at UPM (Madrid, Spain). http://idafensp.github.io/ResearchObjects/WICUS_Makeflow_Blast/ and http://www.sciencedirect.com/science/article/pii/S0167739X16000029

[Szalay 2008] Szalay, AS: *The Sloan Digital Sky Survey and beyond*, SIGMOD Record, June 2008, Vol. 37, No. 2.

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 organisations or persons (e.g. rights holder)
5. 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).

And finally:

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