iCurate: A Research Data Management System

Shuo Liang, Violeta Holmes, Grigoris Antoniou and Joshua Higgins

University of Huddersfield,Huddersfield, United Kingdom
shuo.liang,v.holmes,g.antoniou,joshua.higgins@hud.ac.uk

Abstract. Scientific research activities generate a large amount of data, which varies in format, volume, structure and ownership. Although there are revision control systems and databases developed for data archiving, the traditional data management methods are not suitable for High-Performance Computing (HPC) systems. The files in such systems do not have semantic annotations and cannot be archived and managed for public dissemination.
We have proposed and developed a Research Data Management (RDM) system, iCurate', which provides easy-to-use RDM facilities with semantic annotations. The system incorporates Metadata Retrieval, Departmental Archiving, Workflow Management System, Meta data Validation and Self Inferencing. The 'i' emphasises the user-oriented design. iCurate will support researchers by annotating their data in a clearer and machine readable way from its production to publication for the future reuse.

Keywords: Big Data, Data Curation, Linked Open Data, Research Data Management, Research Support, Semantic Web, Human-Computer Interaction

1 Introduction

At many research institutions, scientific research activities generate a large amount of data, which vary in format, volume, structure and ownership. The data could also be generated on various media. The major challenge in the research community is converting the research data into a generally machine readable format to enable further reuse of the content. Researchers often store their data without proper consideration of security, availability and curation of data. The data includes simulation data in HPC (High-Performance Computing) facilities, raw data from scientific equipment, video and image, text document, etc. Currently, the researchers store their data on various media, often large portable disks. Many research institutions have RDM policies in compliance with the research funding body requirements.

Several tools were suggested by DCC (Digital Curation Centre)[6] to help digital preservation. However, learning methods of preservation and usage of these tools is costly and consumes time. Combining these tools into institutions' existing ways of the research process is also expensive and time consuming.

In the current situation, HPC users' data can not be directly categorised and integrated into the curation process. The HPC file storage system is often maintained at a department level, and not integrated into an institutional storage and cloud storage. The disparate data storage systems are hindering the management process.

Therefore, it is demanded to develop a tool or service in order to improve the current situation of RDM (Research Data Management) and comply with the institutions' RDM policies. However, the research institutions' RDM solutions are not tailored particularly for HPC RDM or enhanced with artificial intelligence to reduce users' effort.

In this paper, we present an overview of current methods of RDM, and evaluate their suitability in the context of the HPC systems. We propose an RDM system, 'iCurate', which provides easy-to-use RDM facilities, combined with state-of-the-art semantic access. We will focus on preservation and curation services of research data, and explore the feasibility of linking research data to related publications maintained within a research institution or university repository. The system aims to integrate multiple storage media and university repository services to serve the digital curation purpose. We have designed a number of components for the existing HPC systems to facilitate easy access across these infrastructures.

2 Background

2.1 Defining Research Data

In a recent public consultation on open research data of the European Commission, the question 'what is research data' was addressed by researchers, funders, industry, etc. In the researchers' point of view,"research data are all data from an experiment, study or measurement, including the metadata and processing details." [4] This reflects that researchers focus more on the data they collected and processed. One view from a funder, The Open Knowledge Foundation, argued that 'research data is extremely heterogeneous and that it takes a variety of forms". [4]

2.2 Ontology and Linked Open Data

One way of increasing data reusability is to publish them with a schema mapped to LOV (Linked Open Vocabularies). LOV is the set of ontologies behind LOD (Linked Open Data). The PREMIS (Preservation Metadata: Implementation Strategies) Ontology is a digital preservation standard based on the OAIS (Open Archival Information System) reference model. It is published as part of LOV by Library of Congress. OAIS is designed to support digital libraries. It is an international standard of archiving. [5] OAIS functional model is composed of six functional entities and related interfaces: Ingest, Archival Storage, Access, Data Management, Preservation Planning and Administration.

2.3 Current and Past Projects for Data Repository

	DataStage	DataBank	DataUp[10]	DMPOnline[3]	Taverna[8]
Metadata Retrieval	✓	✓	✓		✓
Workflow Evaluation				✓	✓
Workflow Management					
Departmental Archiving		✓	✓		
Repository	✓				
OAIS Model					
Metadata Validation					

Table 1. A comparison of current RDM solutions

In table 1, some of the key features of existing RDM tool are compared.

The DataStage and DataBank are departmental data depository tool and University Repository tool developed by DataFlow team. Both tools are open-source. DataStage is designed as a customisable network-attached storage, which could be mounted on end-user's PC as "a mapped drive" for research group level. It has been claimed web-accessible, metadata enrich-able, backup enabled, group shared, password protected, etc. [1] DataBank is a data repository designed for institutions. It provides "a definitive, sustainable, reference-able location for research datasets". [2] DataBank assigns DOIs (Digital Object Identifiers) to datasets in the data repository.

The DataBank projects did not consider the various data storage structure and flexibility for a fragmented storage system. In the application level, DataBank can convert deposited data into publications as a repository. Some other applications such as DataUP and Taverna focus on a specific discipline.

Regardless of compatibility, there are various applications designed for every stage of RDM process. Where a required feature is missing, institutions can combine the applications together, just as the Data Management Toolkits in Penn State University. [9]

However, a whole RDM process will change the researchers' habits. This change may not be suitable to each individual. Furthermore, it affects the quality of the whole RDM plan. The SWORD protocol [7] that DataBank used has set a foundation of a few RDM programs, while other Web service formats such as WSDL and REST have already been used in Digital Preservation area. The co-existence of these formats have made fragmentations in RDM domain.

3 Motivation

The RDM is considered an extension of traditional library databases, thus are built around library repositories because this is the traditional way of publishing. RDM as a part of the research process is preferred ideally to support recording the missing meta-data associated with the data that has been generated or

archived that is related to a future publication. Nevertheless, a standard has not yet been proposed, which defines how detailed the meta-data should be, how to grade the meta-data, and how to improve the process on the scale of time.

The meta-data collection and evaluation tools seen in the previous section are intrusive to a user's own working routine, which makes them difficult to adopt. It is an extra work to manually categorise and group data in a separate system or repository to facilitate annotation, even more so to repeat this process.

Moreover, especially in the HPC field, the computational power is not as fast as the generation rate of the data. The data that has been preserved today needs to be pre-processed before being archived, in order to serve the further re-use. The pre-processed data from an HPC system will have more reusability with rich meta-data.

4 System Design - iCurate

4.1 Main Features

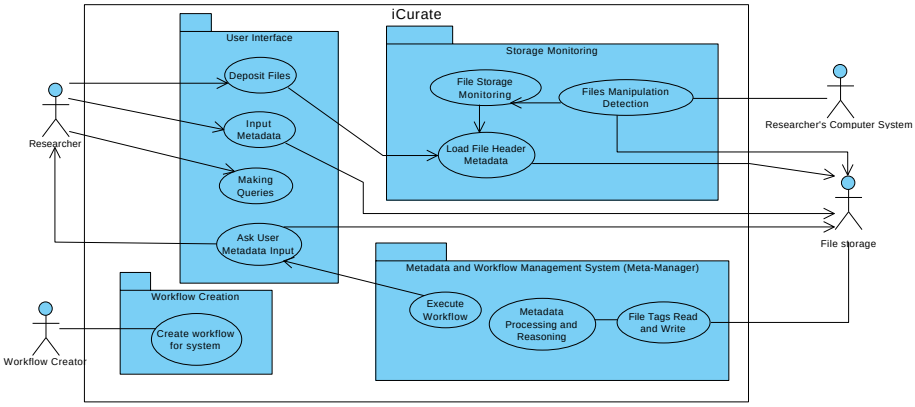

Fig. 1. The use case diagram of iCurate

Based on our investigation outlined in Section 2, it is evident that the existing solutions do not completely satisfy the needs of RDM for users of HPC. The current solutions do not generate and summarise the information of research activities. Hence, we design and develop an RDM system (iCurate) that enables curation process from research data's production to its publication. The 'i' emphasises the user-oriented design philosophy. This system will overcome some limitations of the existing systems and have following features: metadata retrieval, departmental archiving, workflow system, ontology-driven self-inferencing.

1. Metadata Retrieval Collecting Metadata is an essential component of an RDM system. Some meta-data which are generated by particular applications

has already been stored in the header of a file. The file headers form the basic meta-data of iCurate. Furthermore, researchers' other activities such as file manipulation, editing and work schedule are recorded as part of meta-data as well.

2. Departmental Archiving iCurate will be appearing on researcher's computer or HPC facilities as a virtual folder or drive. Researcher's files appears on iCurate in structured view, e.g. tree structure or tag cloud. regardless how they were stored on the drive. iCurate compresses files in the order of curation.

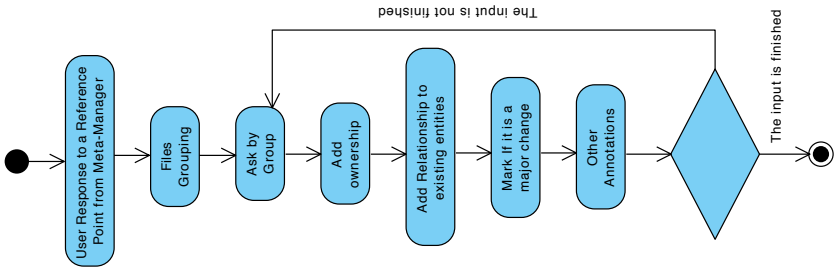

Fig. 2. A use case diagram of iCurate

3. Workflow System In order to overcome the extra contributing efforts in RDM, workflow management is suggested to be added into an RDM system. The workflow system is used in the business process and many computer programs. By sharing workflows of the processing methods which are employed by researchers or anyone who works in similar discipline through the community, people will not be planning and designing their own RDM plan alone, and will contribute more effort on their research or workflow improvement.

4. Metadata Validation and Self Inferencing Being compatible to OAIS and also PREMIS Ontology, will give the RDM system the ability to expose easily not only the public but also Linked Open Data. Inferencing is a step further to workflow management. The inferencing, which is based on the pre-configured workflow and ontologies, gives iCurate error checking, auto-suggestions and other new features. Inferencing requires all meta-data to be mapped onto RDM ontologies to enable the inferencing system. Once the RDM system have collected the data, the workflow system can be programmed to check if there is any contradiction among the data and give suggestions. Users should be able to log-in to a web front-end to perform daily error checking or between any period which are considered appropriate.

In the use case of iCurate, the user interface is depicted in Fig.1. There are two human actors making inputs to the system. A workflow creator defines the workflow of RDM process, it could be adopted by researchers, who are actually recording their metadata. A workflow creator can be a project director, an external user who created and shared his/her RDM workflow, or the researchers

themselves. Once the workflow is defined in a particular user's case, the system extracts meta-data from file storage actor, and generating questionnaires based on the pre-defined workflows, extracts further information about the research work. In the HPC environment it is common for storage to be centralised, such as an external NAS (Network-Attached Storage). This is convenient to monitor files for changes as they are created or modified, unlike existing systems which require the user to manually upload changes to the RDM system.

4.2 Meta-data and Workflow Management

As mentioned above, a workflow creator can be a researcher or another person. In order to cater for particular cases, the workflow can be continuously amended for the new requirement. In Fig.2, an example of a workflow is given. The File Sets are determined by the folder structure and editing time. By grouping files into File Sets, the minimal intellectual entity can be identified. Requested meta-data will be asked after the File Sets are defined by the user.

A database, which collects and stores user inputed meta-data, is convenient but not yet smart. Based on the meta-data set obtained from researchers and reasoning engine, iCurate infers information inconsistency and gives research feedback to correct. Eventually, it will benefit not only the data management but also research activity itself.

4.3 Web Interface

The web interface of iCurate has two functions: to give a summary of the healthiness of researcher's meta-data in the dash board, and to collect meta-data. The dash board gives user an overview of their metadata inputs, project progresses, to-do lists.

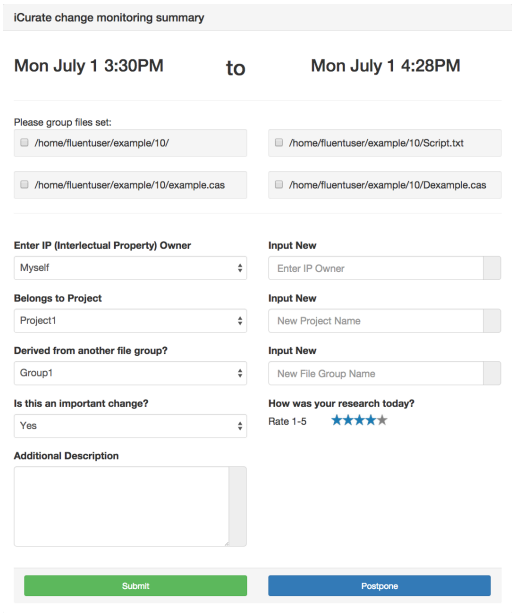

Fig. 3. iCurate Web Change Form

Fig.3 shows a meta-data collection form of iCurate. The user is notified at the time which is predefined by their RDM workflow, e.g. every end of the day, after activity threshold has been reached. The questions are dynamically arranged depending on user's previous inputs. The activity period is highlighted at the top. The files are changed during this period is shown and can be checked to group up as a set of a single entity. The system pre-fill some of the questions based on users' activities and RDM workflow. User will be able to

```
#!/bin/bash
#HPC job related environment setups......
#< . . >

# Exrta infomation for iCurate
IC_INPUT_FOLDER="/home/fluentuser/example/10/"
IC_INPUT_SCRIPT="/home/fluentuser/example/10/Script.txt"
IC_INPUT_FILE="/home/fluentuser/example/10/example.cas"
IC_OUTPUT_FILE="/home/fluentuser/example/10/Dexample.cas"
IC_EXEC="/apps/Fluent/bin/fluent 3ddp -g -env -ssh -mpi=openmpi -t$nprocs \
        -cnf=$PBS_NODEFILE -i $IC_INPUT_SCRIPT"

# Execute command to launch Fluent
$IC_EXEC
```

Fig. 4. An example of modified PBS script

assign the changes to existing projects and File Sets or create new ones. Users can also rate their own research productivity for a reference of their progress, what will be summarised in the dash board.

4.4 HPC Enabled

The iCurate will integrate with PBS (Portable Batch System). The users can declare their meta-data in their job file, and the system-generated meta-data, including successful completion, running time, time started, and so on, will also be recorded. Since every PBS job has its unique number with its hostname, it will be easier to locate the output. However, it is easier to save the screen output from PBS jobs, file outputs can be archived once it is specified. In iCurate, the output files or folder can be specified. Once the jobs are completed, the files or folders will be 'snapshotted' and stored in a separate section.

Fig.4 shows a PBS script with iCurate enabled. More informations are required to be declared before a job submission (Variables name starts with `IC_`). Those Variables and the system record on particular successful submission will be used in iCurate core for tracing user activities.

4.5 Implementation

The implementation of iCurate has three parts, core services, NAS, and User Interface, as shown in Fig.5. The core services can be deployed on the NAS server or an independent server for more stability. Its components include: a Jena RDF (Resource Description Framework) database and reasoning engine, a "Meta-Manager" handling input and output, and a File Storage Monitor extracting user disk activities. The data storage is a modified Linux NAS service, which uses Linux `inotify` to monitor user file system changes. The Web interface is a cross-platform solution for metadata collection. The web interface contains a summary and statistics of user's activities, a score of metadata completion, and metadata questionnaires.

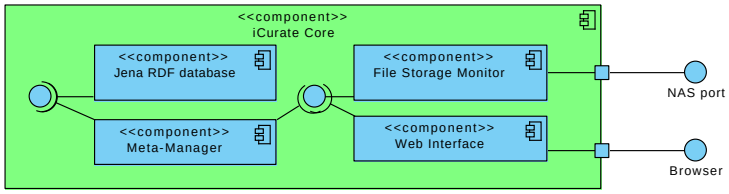

Fig. 5. The components diagram of iCurate

5 Summary

RDM plays a significant part of any research activity and is gaining in importance as a large amount of data is generated from scientific instruments simulations and modelling. In this paper, we have outlined several tools for generation, preservation and curation of data, and identified some of the limitations on the data which are generated in HPC research environments. we have presented out efforts in creating an RDM system, iCurate, which was designed and developed with researchers in mind. This approach allows meta-data be refinedsplncs03 before it is required by formal archiving services. iCurate will provide easy-to-use RDM facility for HPC RDM and it integrates with university repository services. Our future work will involve further development of the system to be deployed not only in the HPC research environment but to be customised for other research domains.

References

1. DataFlow: About datastage, `http://www.dataflow.ox.ac.uk/index.php/datastage/ds-about`
2. DataFlow: Databank, `http://www.dataflow.ox.ac.uk/index.php/databank`
3. (DCC), D.C.C.: Dmponline (2015), `https://dmponline.dcc.ac.uk`
4. European Commission: Results of the consultation on open research data (July 2013), `http://ec.europa.eu/digital-agenda/node/67533`
5. ISO: Iso 14721:2003 subscribe to updates space data and information transfer systems – open archival information system – reference model (2012), `http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683`
6. Jones, S.: New checklist for a data management plan (July 2013), `http://www.dcc.ac.uk/resources/data-management-plans/checklist`
7. Julie Allinson, Sebastien Francois, S.L.: SWORD: Simple Web-service Offering Repository Deposit. Ariadne 54 (January 2008), `http://www.ariadne.ac.uk/issue54/allinson-et-al/`
8. Taverna: Taverna (2014), `http://www.taverna.org.uk/`
9. The Pennsylvania State University: Data management toolkit (Dec 2014), `https://www.libraries.psu.edu/psul/researchguides/pubcur/datatoolkit.html`
10. The Regents of the University of California: Dataup: Describe, manage and share your data, `http://dataup.cdlib.org`