Data Management Plan

E-CAM Deliverable 11.5
Deliverable Type: Report
Delivered in Month 12– September 2016

E-CAM
The European Centre of Excellence for
Software, Training and Consultancy
in Simulation and Modelling

Funded by the European Union under grant agreement 676531

Project and Deliverable Information

Project Title	E-CAM: An e-infrastructure for software, training and discussion in simulation and modelling
Project Ref.	Grant Agreement 676531
Project Website	https://www.e-cam2020.eu
EC Project Officer	Dimitrios Axiotis
Deliverable ID	D11.5
Deliverable Nature	Report
Dissemination Level	Public
Contractual Date of Delivery	Project Month 12(September 2016)
Actual Date of Delivery	28.10.2016

Document Control Information

Document	Title:	Data Management Plan
	ID:	D11.5
	Version:	As of October 28, 2016
	Status:	Draft
	Available at:	https://www.e-cam2020.eu/deliverables
	Document history:	Internal Project Management Link
Review	Review Status:	Not reviewed
	Action Requested:	Resubmit to EU
Authorship	Written by:	Kate Collins(NUID-UCD)
	Contributors:	Alan O'Cais (FZJ-JSC), Jenny O'Neill and Donal Mac Kernan (NUID-UCD)
	Reviewed by:	Reviewer (Institution)
	Approved by:	WPLeader (Institution)

Document Keywords

Keywords:	E-CAM, HPC, CECAM, Materials, ...

October 28, 2016

***Disclaimer**:This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement. It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements.*

[1] kate.collins@ucd.ie

Contents

Executive Summary

E-CAM activities can be divided into three complementary areas and associated types of data object: 1) software and algorithm development; 2) advanced training in the production, documentation and use of scientific software; 3) outreach to industry and academia to identify evolving scientific software needs and opportunities. The objective of the present deliverable is to describe at a high level how we plan to manage the data generated from these activities.

All code developed will follow, where practical, the Extended Software Development Workshop (ESDW) Technical Software Guidelines, and will be subject to the quality acceptance criteria defined by the Guidelines for the ESDW's. Software development will be guided by requests from end-users through the E-CAM website, through industry scoping workshops and through direct collaborative projects with industry. The associated data relating to the project will be maintained in the form of software and metadata repositories (version-controlled through Git).

E-CAM will create software modules rather than complete packages to allow for the rapid inclusion of new algorithmic ideas and their effective dissemination. The project will interface these modules to existing software codes either directly, or through translators. Where a software module is standalone and generated entirely within E-CAM project, the source code software repository will be created and maintained on the E-CAM GitLab service. Where modules relate to externally maintained software packages, appropriate links will be provided in the metadata repository as well as patch files that detail the changes/additions to the source code.

Training material generated directly by the E-CAM project will be made publicly available through a training material repository on our GitLab service following the best practices and guidelines of the Software Carpentry Foundation. The training material will be created using the Software Carpentry approach, meaning that the content will itself be stored in a version-controlled repository.

Our *State of the Art* and and *Industry Scoping* workshops are all required to write reports of their activities. Templates to facilitate community exploitation of data produced through these events have been defined, and are being updated to maximise their impact.

There is also a quarterly newsletter distributed to all project partners and the E-CAM mailing list as well as other reports scheduled for distribution to that list, including information regarding the E-CAM ESDW and industry pilot project outcomes. All of these reports will be stored at EPFL and be accessible through the E-CAM website.

1 Dataset Descriptions

E-CAM activities can be divided into three complementary areas and associated types of data object:

- Software and algorithm development (including testing and documentation) where the resultant data objects are software artefacts, associated data for testing/verification and metadata related to the performance of application software on High Performance Computing (HPC) systems.
- Advanced training in the production, documentation and use of scientific software for use on current and future computational infrastructures. The associated data objects are the training material provided during training events and the documentation of any related identifiable best practice.
- Outreach to industry and academia to identify evolving scientific software needs and opportunities (within the scope of E-CAM). The associated data objects are the emails, newsletters and reports that E-CAM will provide to it's industrial and academic user communities over the project lifetime.

The objective of the present deliverable is to describe at a high level how we plan to manage the data generated from these activities. The outline provided here is to ensure that generated data is stored in a coherent, accessible and reusable fashion; that data exchange between members of the consortium (and the wider scientific community where relevant) is both technically straightforward and adequately supported; and that a long-term data access and archiving plan is in place.

1.1 Software and Associated Metadata

One of the core aims of the E-CAM project is to create some 150 new robust software modules which will describe, interface with and/or extend existing standard packages as well as being a library of software generated directly within E-CAM. These software modules will span E-CAM's four scientific areas of:

- classical molecular dynamics,
- electronic structure,
- quantum dynamics,
- and meso and multi-scale modeling.

Software development will be guided by requests from end-users through the E-CAM website, through industry scoping workshops and through direct collaborative projects with industry. The associated data relating to the project will be maintained in the form of software and metadata repositories (version-controlled through Git).

1.1.1 Guidelines for Asset Creation

Where practical, all code developed will follow the ESDW Technical Software Guidelines and will be subject to the quality acceptance criteria defined by the Guidelines for the ESDW's (which are updated annually).

The specification, documentation and verification of the development of the module is considered metadata and this will form part of the metadata repositories of E-CAM. The metadata repositories will be directly maintained and managed by E-CAM and will contain the main documentation infrastructure of E-CAM (referred to in the original proposal as the E-CAM "wiki").

1.1.2 Software Assets

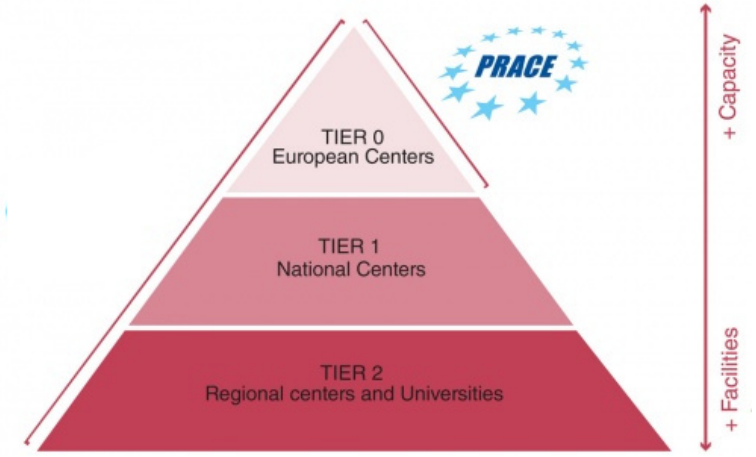
E-CAM will create software modules, rather than complete packages, to allow for the rapid inclusion of new algorithmic ideas and their effective dissemination. The project will interface these modules to existing software codes either directly or through translators (specific software tools that convert the output of one package to the input of another).

Where a software module is standalone and generated entirely within E-CAM project, the source code software repository will be created and maintained on the E-CAM GitLab service. Where modules relate to externally maintained software packages, appropriate links will be provided in the metadata repository as well as patch files that detail the changes/additions to the source code (however these can only be provided subject to licensing).

1.1.3 Metadata Assets

We currently foresee two distinct metadata repositories:

- an application repository storing information on software applications used by the E-CAM community, including build and performance information on HPC architectures. The provided build information will be tuned to work efficiently on the top three tiers of the hardware pyramid (PRACE, 2008, p.13).



- an E-CAM module repository that contains detailed documentation on the modules developed within the E-CAM ESDWs and by the E-CAM post-docs. This repository will include information on software development best practices (both in general and specific to HPC systems), forming the "living" version of the ESDW Technical Software Guidelines.

These metadata repository will be constructed in such a way that it can viewed as a wiki and built as a stand-alone document using the Sphinx documentation tool. The wiki version of the repositories will be delivered through readthedocs.org.

1.2 Advanced Training

Where possible we will use and/or provide open source training material. We will provide appropriate links on the training section of the E-CAM website. When training is provided by a third party we will attempt to ensure that this content is also made publicly available.

1.2.1 Training Assets

Training material generated directly by the E-CAM project will be made publicly available through a training material repository on our GitLab service following the best practices and guidelines of the Software Carpentry Foundation. The training material will be created using the Software Carpentry approach, meaning that the content will itself be stored in a version-controlled repository. Examples of the Software Carpentry lessons as raw repositories and rendered websites can be found on the Software Carpentry lesson pages.

Where training is provided by third parties, training content (such as documents, captured presentations, examples, etc.) will, subject to obtaining their permission, be uploaded or linked to the training section of the E-CAM website.

1.3 Outreach to Industry and Academia

Our *State of the Art* and and *Industry Scoping* workshops are all required to write reports of their activities. Templates to facilitate community exploitation of data produced through these events have been defined, and are being updated to maximise their impact.

There is also a quarterly newsletter distributed to all project partners and the E-CAM mailing list as well as other reports scheduled for distribution to that list, including information regarding the E-CAM ESDW and industry pilot project outcomes.

All of these reports will be stored at EPFL and be accessible through the E-CAM website.

2 Data Access and Sharing

All software and metadata repositories directly generated by E-CAM will be open source and accessible via the E-CAM website. These are public repositories based in our GitLab service.The GitLab repositories will be updated continuously throughout the project lifetime and will remain after the project lifetime. Access information will be publicly posted and advertised in publications. There will be no charge for accessing the libraries.

The software will be open source, highly structured and well-documented in line with the Pilot on Open Research Data as specified in the Model Grant Agreement (Article 29.3). Doxygen will be used to document source code (where no alternative is already in use). All repositories will also have complete documentation (based on ReStructured Text and Sphinx) and their online version will be hosted on readthedocs.org. The documentation itself will be part of the E-CAM repositories so will be versioned and version-controlled in the same manner as source code. ReStructured Text is also rendered by GitLab, so a large portion of documentation will be directly readable from the GitLab repositories.

The libraries do not contain 'personal data' and personal information is not required to access the libraries. We will provide a form on the E-CAM website where people can provide contact information to be updated on developments within the project. This data will be stored within a secure database, used internally by the project and will not be shared with third parties.

2.1 Licencing

All software will be licensed for re-use and re-distribution, the specific restrictions of this is dependent on the particular licence of the particular software (and the project cannot dictate the licence terms for external software used). Open-source licences such as GPL, LGPL or FreeBSD will be used where possible. It is expected that the software will be used primarily by researchers in academia and industry working in the field of simulation and modelling. The intended or foreseeable uses / users of the data would be those seeking to improve the efficiency, performance or capability of simulation software and there are no reasons not to unnecessarily restrict data sharing or re-use.

The licensing for non-software assets (such as wiki entries, scaling plots etc.) generated by E-CAM will use the Creative Commons Attribution 4.0 International License.

2.2 Referencing

Each module created by E-CAM (as a documented software asset) will be referenceable by direct URL to readthedocs.org.

Metadata assets, since they are documented with ReStructured Text and readthedocs.org, will also be referenceable in a similar manner.

We encourage the E-CAM community to create specific releases of their software products and upload these releases to services such as Zenodo in order to ensure that their software applications are citable.

2.3 Right of Management

E-CAM maintains administration rights over the E-CAM GitLab service. Within the service it is possible to manage repositories within groups. Groups exist for the E-CAM organisation and for each research Work Package (WP). E-CAM will maintain the rights of management over these groups and these rights shall pass to Centre Européen de Calcul Atomique et Moléculaire (CECAM) should E-CAM be wound down.

Rights of management for other repositories within the service shall be maintained by the respective owners.

Any third party may request additions, corrections or similar to any of the repositories via the use of GitLab "Issues" and "Merge Requests"

2.4 Intellectual Property Rights

Intellectual property rights will adhere to the terms of the E-CAM Grant Agreement with the European Commission, and the E-CAM Consortium agreement. As not all intellectual property developed within the E-CAM initiative will be generated solely by authors employed by E-CAM beneficiaries, intellectual property will remain the property of it's creator(s), or the employer(s) of its creator(s) if the property was created under the normal course of their employment (unless a contractual agreement exists stating otherwise).

2.5 E-CAM Website

The reference contact for the community outside E-CAM is our E-CAM website. Here are defined the purposes and the people involved in the project,resources (like the link to our GitLab service mentioned above), deliverables and Quarterly Newsletter updates.

2.5.1 Internal data sharing

The management and organisation of all the documents, deliverable, etc. relevant to the E-CAM project is done by using the web-based software manager Redmine. For each WP a dedicate section has been created. It contains a brief description of the package, the planned Tasks, Milestones, Deliverables, Meetings, Subtasks, Support and Quarterly Reports.

3 Archiving and Preservation

There is a long-term strategy for maintaining, curating and archiving the software repositories through the data-office located at EPFL and through our GitLab service. Both the application library and the software library will be versioned. As such, the software will be usable beyond the project life-time and will be placed in Zenodo, the data repository of the OpenAIRE (Open Access Infrastructure for Research in Europe) project as recommended by the Pilot on Open Research Data.

The application and software libraries are either source code or plain text (with some images for examples and some testing data). In terms of storage capacity, 150 modules is unlikely to take more than a 1GB of storage capacity. The releases of each library will be archived on the archive facilities at the data offices at EPFL, but these will not be publicly accessible. By it's nature, contributing to the GitLab repositories creates a distribution of copies of the repositories with a complete contribution history. We will also mirror our repositories on GitHub. In addition to this we will archive our releases in the archive facilities of EPFL.

Using `git`, GitLab and mirroring our public repositories on GitHub means our software is searchable, public and easily accessible using what will be a familiar interface for many people. In addition, anyone with a GitHub account will be able to make contributions to either library and we can take advantage of the additional services that can link with GitHub (such as Zenodo).

3.1 Responsibility

Reponsibility for the preservation of the GitLab service, it's repositories, associated meta-data and back-ups of same lie with CECAM, and specifically through EPFL.

References

Acronyms Used

CECAM Centre Européen de Calcul Atomique et Moléculaire

HPC High Performance Computing

ESDW Extended Software Development Workshop

WP Work Package

URLs referenced

Page ii

https://www.e-cam2020.eu ... https://www.e-cam2020.eu
https://www.e-cam2020.eu/deliverables ... https://www.e-cam2020.eu/deliverables
Internal Project Management Link ... https://redmine.e-cam2020.eu/issues/116
kate.collins@ucd.ie ... mailto:kate.collins@ucd.ie
http://creativecommons.org/licenses/by/4.0 ... http://creativecommons.org/licenses/by/4.0

Page 1

ESDW Technical Software Guidelines ... https://www.e-cam2020.eu/wp-content/uploads/2016/05/D6.1-Guidelines-for-web-1.pdf
Guidelines for the ESDW's ... https://www.e-cam2020.eu/wp-content/uploads/2016/05/Deliverable-5.1-final.pdf
Git ... https://git-scm.com/
E-CAM GitLab service ... https://gitlab.e-cam2020.eu/explore/projects
patch files ... https://en.wikipedia.org/wiki/Patch_(Unix)#Patches_in_software_development
our GitLab service ... https://gitlab.e-cam2020.eu/explore/projects
Software Carpentry Foundation ... http://software-carpentry.org/

Page 2

Git ... https://git-scm.com/
ESDW Technical Software Guidelines ... https://www.e-cam2020.eu/wp-content/uploads/2016/05/D6.1-Guidelines-for-web-1.pdf
Guidelines for the ESDW's ... https://www.e-cam2020.eu/wp-content/uploads/2016/05/Deliverable-5.1-final.pdf
E-CAM GitLab service ... https://gitlab.e-cam2020.eu/explore/projects
patch files ... https://en.wikipedia.org/wiki/Patch_(Unix)#Patches_in_software_development

Page 3

ESDW Technical Software Guidelines ... https://www.e-cam2020.eu/wp-content/uploads/2016/05/D6.1-Guidelines-for-web-1.pdf
Sphinx documentation tool ... http://www.sphinx-doc.org/en/stable/
readthedocs.org ... https://readthedocs.org/
training section of the E-CAM website ... https://www.e-cam2020.eu/training/
our GitLab service ... https://gitlab.e-cam2020.eu/explore/projects
Software Carpentry Foundation ... http://software-carpentry.org/
Software Carpentry lesson pages ... http://software-carpentry.org/lessons/
training section of the E-CAM website ... https://www.e-cam2020.eu/training/

Page 4

our GitLab service ... https://gitlab.e-cam2020.eu/explore/projects
Doxygen ... http://www.stack.nl/~dimitri/doxygen/
Sphinx ... http://www.sphinx-doc.org/en/stable/
readthedocs.org ... https://readthedocs.org/
readthedocs.org ... https://readthedocs.org/
readthedocs.org ... https://readthedocs.org/
Zenodo ... https://zenodo.org/
E-CAM GitLab service ... https://gitlab.e-cam2020.eu/explore/projects
"Issues" ... https://docs.gitlab.com/ee/gitlab-basics/create-issue.html
"Merge Requests" ... https://docs.gitlab.com/ce/workflow/forking_workflow.html

Page 5

E-CAM website . . . http://https://www.e-cam2020.eu/
Redmine . . . http://www.redmine.org/

Page 6

our GitLab service . . . https://gitlab.e-cam2020.eu/explore/projects
GitHub . . . https://github.com/
GitLab . . . https://gitlab.e-cam2020.eu/explore/projects
GitHub . . . https://github.com/
Zenodo . . . https://zenodo.org/

Citations