E-CAM Software Platform III

E-CAM Deliverable 6.5
Deliverable Type: Report
Delivered in February, 2018

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	Juan Pelegrín
Deliverable ID	D6.5
Deliverable Nature	Report
Dissemination Level	Public
Contractual Date of Delivery	Project Month 27(31st December, 2017)
Actual Date of Delivery	7th February, 2018
Description of Deliverable	Update on E-CAM software tools and platforms jointly by CH and Juelich which includes: - the E-CAM library of software modules and interfaces; - end users portal (to access E-CAM's resources, make requests for software developments, register for events); - web infrastructure for teaching tools.

Document Control Information

Document	Title:	E-CAM Software Platform III
	ID:	D6.5
	Version:	As of February, 2018
	Status:	Accepted by WP leader
	Available at:	https://www.e-cam2020.eu/deliverables
	Document history:	Internal Project Management Link
Review	Review Status:	Reviewed
	Action Requested:	Submit
Authorship	Written by:	Alan O'Cais(Juelich Supercomputing Centre)
	Contributors:	Ana Mendonça (EPFL)
	Reviewed by:	David Swenson (University of Amsterdam, NL)
	Approved by:	Alan O'Cais (Juelich Supercomputing Centre)

Document Keywords

Keywords:	E-CAM, HPC, CECAM, Materials, Training, Online learning, Portal, Version Control, Best Practices...

7th February, 2018

***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.*

***Copyright notices**: This deliverable was co-ordinated by Alan O'Cais[1] (Juelich Supercomputing Centre) on behalf of the E-CAM consortium with contributions from Ana Mendonça (EPFL) . This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0.*

[1] a.ocais@fz-juelich.de

Contents

Executive Summary **1**

1 Introduction **3**
1.1 Scope of the update . . . 3
1.2 Target audience . . . 3
1.3 Infrastructure delivery and back-up . . . 3

2 Library of Software Modules **4**
2.1 What is a software module? . . . 4
2.2 How are software modules contributed to the library? . . . 5

3 End Users Portal **6**
3.1 Project website . . . 6
3.1.1 Pilot Projects . . . 6
3.2 GitLab . . . 6
3.2.1 Version Control . . . 7
3.2.2 Continuous Integration . . . 7
3.2.3 GitLab Pages . . . 7
3.2.4 Authentication . . . 8
3.3 Project Management . . . 8
3.3.1 Redmine . . . 8
3.3.2 Kanban . . . 8
3.3.3 EtherPad . . . 9
3.3.4 ShareLatex . . . 9

4 Web Infrastructure for Teaching Tools **11**
4.1 Lecture Capture . . . 11
4.1.1 Remote participation . . . 11
4.1.2 Example of captured content . . . 11
4.1.3 Capture of E-CAM events . . . 12
4.2 Content Management . . . 12
4.2.1 Outsourcing . . . 12
4.3 Structuring the Training Portal . . . 14
4.3.1 Datasets . . . 14
4.3.2 Collections . . . 17
4.3.3 Spaces . . . 18
4.4 Future Development . . . 19
4.4.1 Collaboration with other projects . . . 19

References **20**

List of Figures

1 The landing page for the E-CAM Software Library . . . 4
2 A list of the E-CAM repositories on our GitLab service . . . 7
3 A view of the E-CAM project within our Redmine service . . . 8
4 An example use case for our Kanban service . . . 9
5 An example use case for our Etherpad service . . . 9
6 An example use case for our ShareLatex service . . . 10
7 A captured lecture and YouTube playlist of an entire event. . . . 11
8 Captured presentation with navigation carousel. . . . 13
9 Example of a URL in Clowder, in this case the self-assessment form for an Extended Software Development Workshop (ESDW). . . . 13
10 Schema of how we will organise the training material generated by E-CAM in the Clowder infrastructure. 14
11 A PDF file rendered within Clowder. . . . 15
12 A repo provided as a link to GitLab. . . . 15
13 An example of lecture dataset which includes a directory for supporting material (in this case the Git repository of Fig. 12). . . . 16
14 A set of lectures are gathered together into a collection. . . . 17

15 A child collection created from the set of lectures relating to software development within the ESDW. . . 18
16 An example of set of spaces. The "ESDW6" space is for attendees to the associated ESDW, its contents can be transferred to the "E-CAM" space (which has many more members) once the workshop is fully completed. 19

List of Tables

1 E-CAM documentation repositories 4

Executive Summary

This deliverable describes the provision of online services in the E-CAM project which together form the E-CAM web platform. The primary landing point for information about the resources of the project is the E-CAM project website. This site alone covers the basic requirements of the E-CAM User Portal:

- *E-CAM library of software modules and interfaces*

  The software modules of E-CAM are linked through the website to the E-CAM GitLab service where they are hosted.

- *Access to E-CAM's resources*

  All of E-CAM resources are described and available through the E-CAM website. This includes our upcoming E-CAM events and the E-CAM online services.

- *Make requests for software developments*

  We deal with development requests directly on a case-by-case basis with the relevant Work Package (WP) leader and the Software Manager being in direct contact with the person making the request. For this reason we have created a very simple technical first contact page in order to channel users to the correct WP. We have also created a follow-up questionnaire that is modelled on the Type C PRACE Preparatory Access form.

- *Register for events*

  All E-CAM events are funded and managed through CECAM with registration for events happening through them. On the E-CAM website we provide detailed descriptions of the E-CAM events and links to the registration process of CECAM.

With respect to the web infrastructure for teaching tools that E-CAM is developing, significant progress has been made. As regards content, we have a capture and remote participation methodology in place for our training events. This methodology has been successfully demonstrated for the 2nd EasyBuild User Meeting that E-CAM co-organised. Relevant portions of the Extended Software Development Workshop (ESDW) events of 2017 have also been captured.

The outsourced development relating to the E-CAM Training Portal is now complete and the service has been configured and publicly available since December 2017. TThe system upon which our service will be built is Clowder, which is developed at NCSA. Clowder is a research data management system designed to support any data format and multiple research domains. E-CAM has expanded the capabilities of Clowder to understand new filetypes that cover the scope of the content to be includeed by E-CAM: video presentations, PDFs and URLs. Content ofall types is being added from the 2017 ESDW program and the service is expected to be widely leveraged in the 2018 program.

E-CAM has delivered a number of additional online resources and capabilities that go beyond the original scope of this deliverable but are also included and described:

- Software modules are contributed to E-CAM through the documentation repository of the relevant research-related Work Package. The sources for the documentation are stored on the E-CAM GitLab service with rendered documentation available through ReadTheDocs.org.
- A Kanban service has been made available to facilitate a lower setup overhead and direct interaction with the issue reporting features of GitLab.
- The Redmine service is used to manage larger software projects and track related issues. It allows users to manage multiple projects and associated sub-projects. It features project wikis and forums, issue tracking, time tracking, and flexible, role-based access control.
- An Etherpad service has been provided for a number of participants to simultaneously add to meeting notes and minutes during an online collaborative meeting.
- A ShareLatex service has been added to facilitate the collaborative production of publication-quality papers using LaTeX.

Over the lifetime of the project these online services continue to mature and expand, particularly in the case of online learning. We will soon be able to evaluate the effort associated with content creation and the impact of the end product.

With respect to content creation through collaboration with other projects, we are currently collaborating with Partnership for Advanced Computing in Europe (PRACE) on some of the training courses in it's PRACE Advanced Training Centre courses. In particular, we collaborate with JSC in 2018 on an upcoming "High-performance scientific computing in C++" course and a "Python in High Performance computing course". Furthermore we are contact with the Molecular Sciences Software Institute (MolSSI) which is a US-based project similar in scope to E-CAM (but significantly larger).

1 Introduction

In its role as an e-infrastructure for the wider E-CAM community, a number of services have been made available to facilitate the creation of, collaboration on and publication of software projects.

This report is an update to E-CAM Software Platform II that highlights the services that are currently available to E-CAM users and outlines the purpose of each service. These services include

- Online software repositories and associated services,
- A number of project management services to facilitate effective collaboration on software projects and to streamline the creation of associated publications,
- An online training platform populated with the training content delivered at the project's ESDW events.

Over the lifetime of the project these online services will continue to mature and expand, particularly in the case of online learning as new content is added and the platform is increasingly leveraged by training events.

1.1 Scope of the update

Each of the deliverables in this sequence is intended to be self-contained. In this respect a significant portion of what is included here has already been covered in previous iterations of the *E-CAM Software Platform* set of deliverables [2]. As well as an overall update of the document, the primary addition in this deliverable is the delivery of our online learning platform including:

- Completion of related outsourced work,
- Configuration of the Clowder service,
- Creation of an authentication plugin for Clowder that can leverage the authentication service provided by GitLab[3],
- Provision of all related services as Docker[4] containers,
- Integration of an initial ESDW to provide an example workflow.

Full details are in Section 4.

1.2 Target audience

This deliverable documents the status and plans for the software platform that E-CAM is creating for it's target community. As a reporting and planning document, the primary intended audience is our funding body. This document is of limited interest to our wider target community since the services we currently have available are already thoroughly documented on the E-CAM project website (to which they are much more likely to be exposed).

1.3 Infrastructure delivery and back-up

Each E-CAM service is delivered within a Docker container and is run on a special purpose server at Centre Européen de Calcul Atomique et Moléculaire (CECAM) headquarters in Lausanne. This approach makes the backup of E-CAM services efficient and straightforward. It also facilitates easy migration of services to either a new server or a new site in the future.

The disk which stores the services is mirrored (to mitigate the impact of a hardware failure) and the services themselves are remotely backed up (with daily/weekly/monthly versions available).

[2] See [1] and E-CAM Software Platform II
[3] The Clowder authentication plugin for GitLab has been contributed back to the main Clowder repository
[4] Docker is an open-source project that automates the deployment of Linux applications inside software containers.

2 Library of Software Modules

One of the primary outputs of E-CAM are the software modules produced by the postdoctoral researchers of the project and the participants of E-CAM ESDW events.

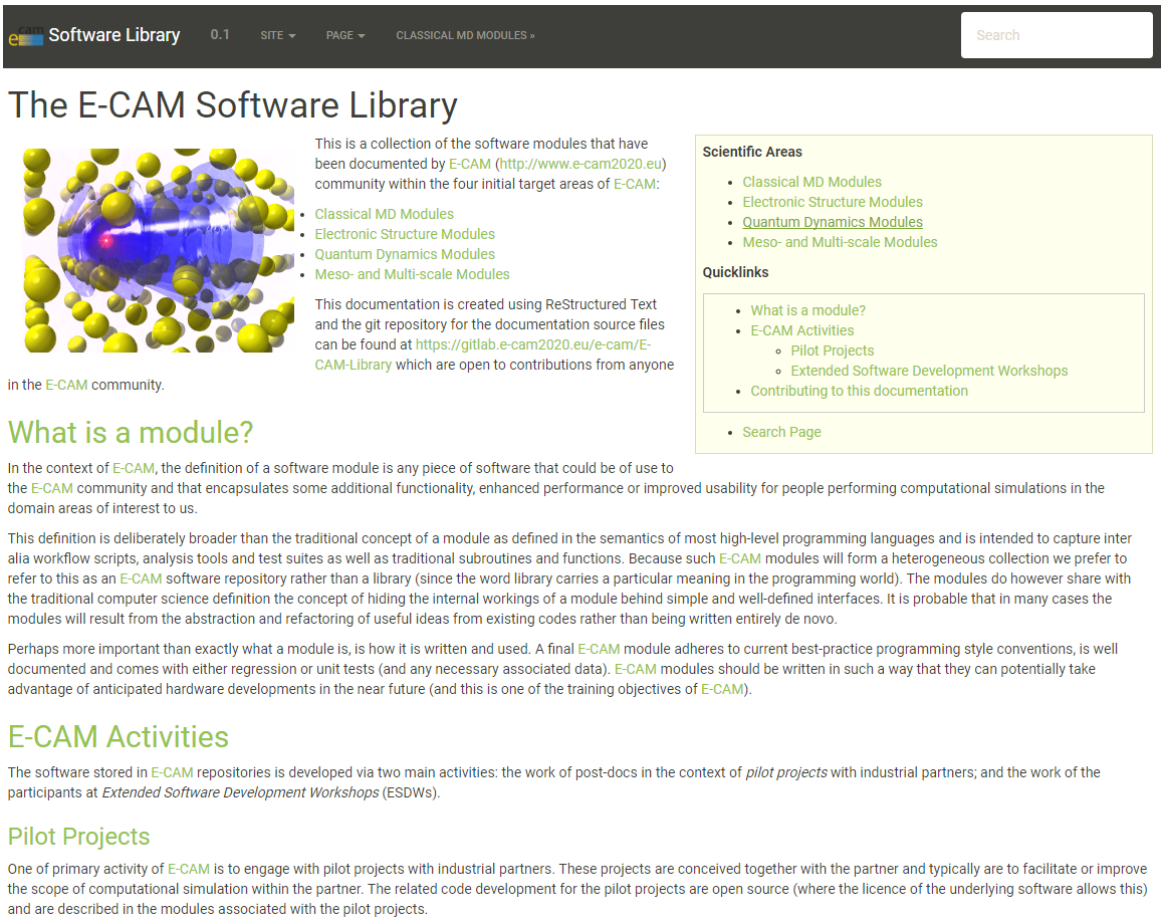

Figure 1: The landing page for the E-CAM Software Library

Initially, software modules are contributed to E-CAM through the documentation repository of the relevant research-related WP (which are shown in Table 1). Currently, all WPs contribute to a combined documention repository: the E-CAM Software Library[5] which is shown in Fig. 1. This allows us to include (and cross-reference) software developments that do not map exclusively to a single research WP, as well as concentrate the development effort associated with the library infrastructure.

WP	*Deprecated webpage*	*Current webpage*
Classical MD	WP1 Deprecated Site	WP1 Current Webpage
Electronic Structure	WP2 Deprecated Site	WP2 Current Webpage
Quantum Dynamics	WP3 Deprecated Site	WP3 Current Webpage
Meso- and Multi-scale Modelling	WP4 Deprecated Site	WP4 Current Webpage

Table 1: E-CAM documentation repositories

2.1 What is a software module?

In the context of E-CAM, the definition of a software module is any piece of software that could be of use to the E-CAM community and that encapsulates some additional functionality, enhanced performance or improved usability for people performing computational simulations in the domain areas of interest to us.

This definition is deliberately broader than the traditional concept of a module as defined in the semantics of most high-level programming languages and is intended to capture inter alia workflow scripts, analysis tools and test suites

[5]The Git repository for this page can be found at https://gitlab.e-cam2020.eu/e-cam/E-CAM-Library

as well as traditional subroutines and functions.

E-CAM modules necessarily form a heterogeneous collection but modules do however share the traditional computer science concept of hiding the internal workings of a module behind simple and well-defined interfaces. It is probable that in many cases the modules will result from the abstraction and refactoring of useful ideas from existing codes rather than being written entirely de novo.

The E-CAM library is a *documentation* repository of software development efforts related to the project. Each module references the source code changes to which it applies, and provides detailed information on the relevant application for the changes as well as how to build and test the associated software.

Perhaps more important than exactly what a module is, is how it is written and used. A final E-CAM module adheres to current best-practice programming style conventions, is well documented and comes with either regression or unit tests (and any necessary associated data). E-CAM modules should be written in such a way that they can potentially take advantage of anticipated hardware developments in the near future (and this is one of the training objectives of E-CAM). Therefore, for a module to be accepted into the E-CAM repository the source code changes in the target application must pass a number of acceptance criteria:

- Style (use meaningful variable names, no global variables,...)
- Source code documentation (each function should be documented with each argument explained)
- Tests (everything you add should have either unit or regression tests)
- Performance (if what you introduce has a significant computational load you should make some performance optimisation effort using an appropriate tool. You should be able to verify that your changes have not introduced unnecessary performance penalties, are threadsafe if needed,...)

2.2 How are software modules contributed to the library?

The library uses the Sphinx documentation generator and the documentation sources are stored on the E-CAM GitLab service (see Section 3.2). Contributions to the repositories are made through *Merge Requests*[6]. Each individual modification of the repository automatically causes the associated documentation on ReadTheDocs.org to be automatically rebuilt.

Within the library itself, E-CAM maintains a set of Module Contribution Guidelines which explains the technical details of submitting a contribution.

[6] Merge (or pull) requests are created in a git management application and ask an assigned person to merge two branches. Tools such as GitHub and Bitbucket choose the name pull request since the first manual action would be to pull the feature branch. Tools such as GitLab and Gitorious choose the name merge request since that is the final action that is requested of the assignee.

3 End Users Portal

3.1 Project website

The primary landing point for access to different information about the project is the E-CAM project website. This site alone covers the basic requirements of the End Users Portal:

- *Access E-CAM's resources*

  All of E-CAM resources are described and available through the E-CAM website. This includes:

  - E-CAM events calendar
  - E-CAM online services
  - Submitted deliverables
  - E-CAM publications and open access information
  - Workshops scientific reports
  - Software repositories for each of the scientific WPs
  - Pilot projects with industry (see section 3.1.1)
  - Job opportunities
  - Latest news items

- *Make requests for software developments*

  Software development request are treated directly on a case-by-case basis with the relevant WP leader and the Software Manager being in direct contact with the person making the request. For this reason we have created a very simple technical first contact page in order to channel users to the correct WP. We have also created a follow-up questionnaire that is modelled on the Type C PRACE Preparatory Access form.

- *Register for events*

  All E-CAM events are funded and managed through CECAM with registration for events happening through them. On the E-CAM website we provide detailed descriptions of the E-CAM events and links to the registration process of CECAM.

The scope of the services offered extend beyond the topics above and E-CAM has delivered a number of additional online resources and capabilities that are described in sections 3.2 and 3.3.

3.1.1 Pilot Projects

The project supports academic and industrial research via a set of pilot projects focused on industrially oriented problems, that are sustained by E-CAM funded Postdoctoral Research Associates (PDRAs) supervised by scientists in the team. Each of these pilot projects has a webpage dedicated to them on the E-CAM website, where the description of the work, modules produced and published results can be seen from people external to E-CAM. Outreach material produced with the results from these pilot projects, including use cases, success stories and interviews, will also be published on this page.

We believe that by doing so, we can increase public visibility towards these projects and E-CAM in particular, and most importantly increase interest by current and potential industrial partners.

The software developed by the PDRAs on the industrial pilot projects will be seen from this page, with a status, expected delivery date for those that are work in progress, and a link to the E-CAM repository where more information about the module documentation and testing can be found. By disseminating the PDRAs work, we also aim at raising employers interest by their professional profile and skills.

3.2 GitLab

We have chosen to create an E-CAM GitLab service as our online repository management system. This allows us the possibility of creating unlimited private repositories for our user community, something that is essential when considering either licensing or the Intellectual Property (IP) requirements of our userbase (particularly those in industry

but also those in academia who may wish to embargo the release of their software until associated results have been published).

As well as being our online Git repository manager, it also provide us with additional repository services such as a wiki, issue tracking, continuous integration, continuous deployment and the ability to create static websites for any GitLab projects. We outline some of these services in the following subsections (and they have also been mentioned in *D6.2: E-CAM software development tools*).

3.2.1 Version Control

The primary software development tool that E-CAM provides is a version-control system. Git is ubiquitous in modern software development and is the obvious choice as the main version control system.

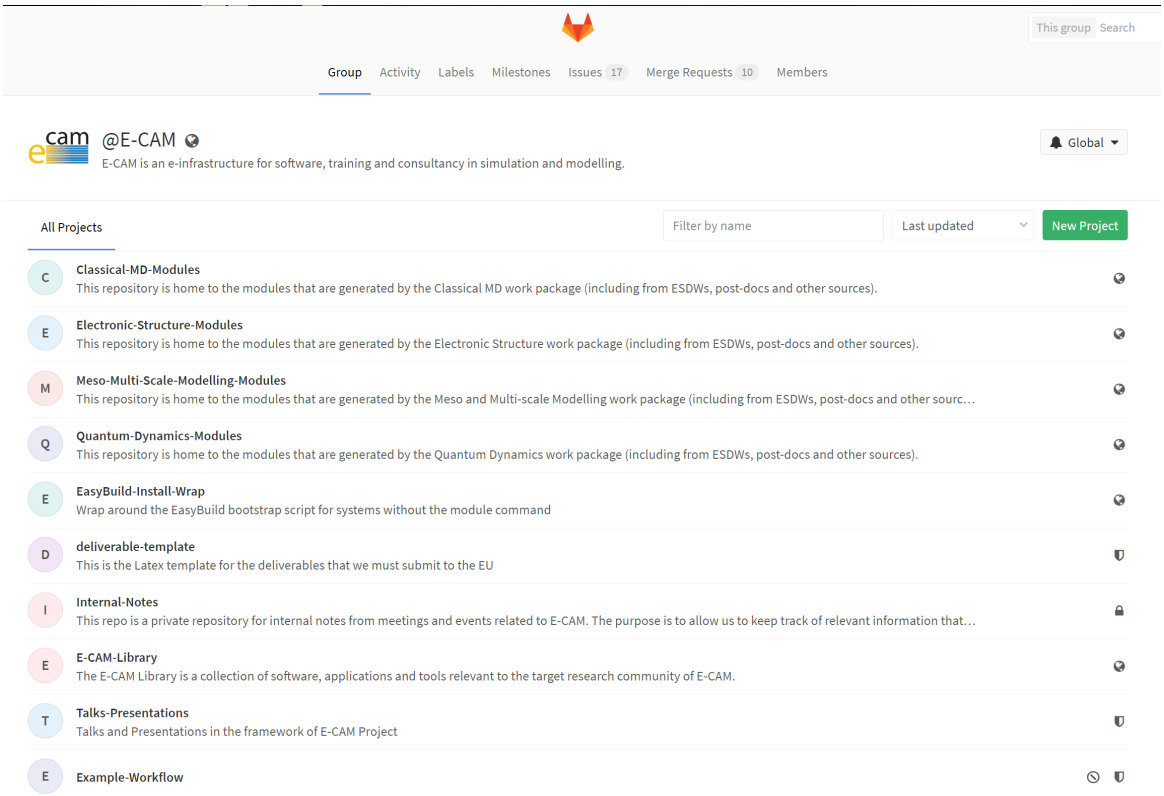

Figure 2: A list of the E-CAM repositories on our GitLab service

A screen shot of the direct repositories of the E-CAM software group can be seen in Fig. 2. Some of these repositories have already been discussed in Section 2. The actual number of software repositories is far greater than this: there are currently 96 users with a total of 177 projects (as of Sept. 2017). On our main project website we also provide a link to all publicly accessible repositories of E-CAM.

3.2.2 Continuous Integration

Continuous Integration is a software development practice in which a developer builds and tests software every time that they create new code in an application. GitLab provides a method of implementing continuous integration and we in turn make it available to our community.

3.2.3 GitLab Pages

With GitLab Pages you can create static websites for your GitLab projects, groups, or user accounts. You can use any static website generator, such as Jekyll, Middleman, Hexo, Hugo, Pelican,... The feature leverages the continuous integration capabilities of GitLab to ensure continuous delivery of the project documentation.

The final webpages are served under the domain `e-cam2020.io` and an example of such a project can be seen for a clone of E-CAM documentation pages.

3.2.4 Authentication

Where possible we leverage the OAuth[7] service of GitLab in order to have single-sign-on capabilities for all E-CAM services. All services except

3.3 Project Management

Supporting and streamlining the management of software projects, including bug-tracking and feature requests, is another area where E-CAM can be of service to the community. Many of these capabilities already exist in our GitLab service, however we have enabled a number of additional services that seek to make remote collaboration on software projects as straightforward as possible.

3.3.1 Redmine

Redmine is a free and open source, web-based project management and issue tracking tool. It allows users to manage multiple projects and associated sub-projects. It features project wikis and forums, time tracking, and flexible, role-based access control. It includes a calendar and Gantt charts to aid visual representation of projects and their deadlines. Redmine integrates with various version control systems and includes a repository browser and a diff viewer.

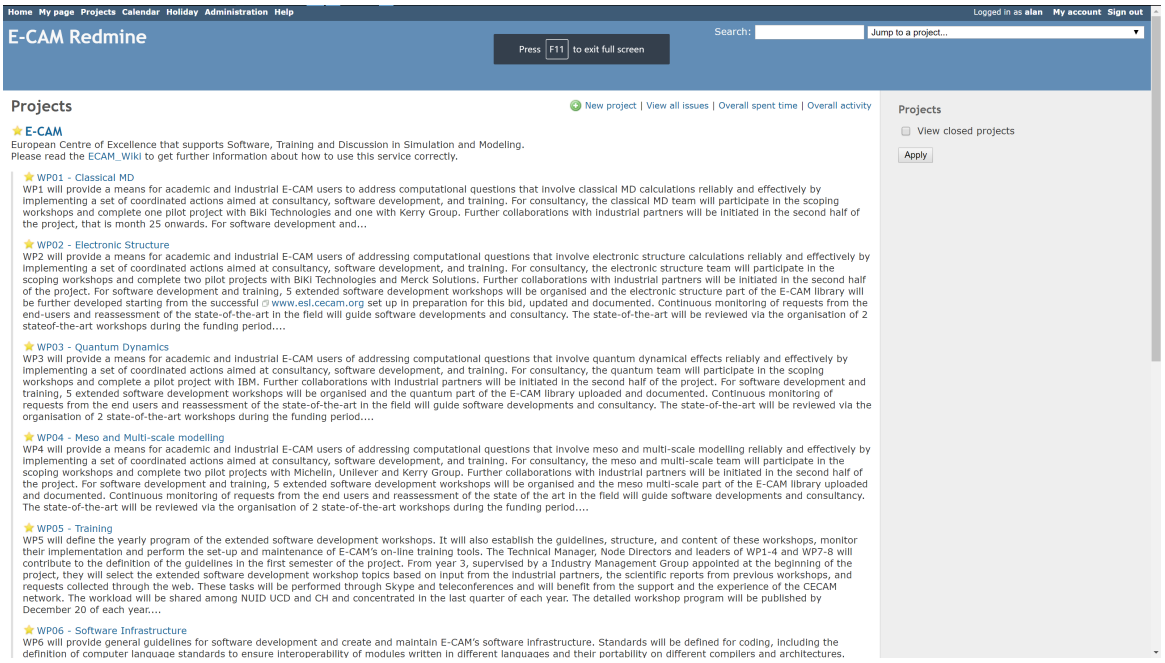

Figure 3: A view of the E-CAM project within our Redmine service

We also use the Redmine service to manage the overall E-CAM project and this example of Redmine usage can be seen in Fig. 3.

3.3.2 Kanban

In order to facilitate a lower setup overhead and direct interaction with the issue reporting features of GitLab, we have also made a Kanban[8] service available that connects directly to the GitLab repositories.

An example of it's use in the case of one of the E-CAM repositories can be seen in Fig. 4

[7] OAuth is an open standard for authorization, commonly used as a way for Internet users to authorize websites or applications to access their information on other websites but without giving them the passwords.

[8] Kanban in the context of software development provides a visual process-management system that aids decision-making about what to produce, when to produce it, and how much to produce.

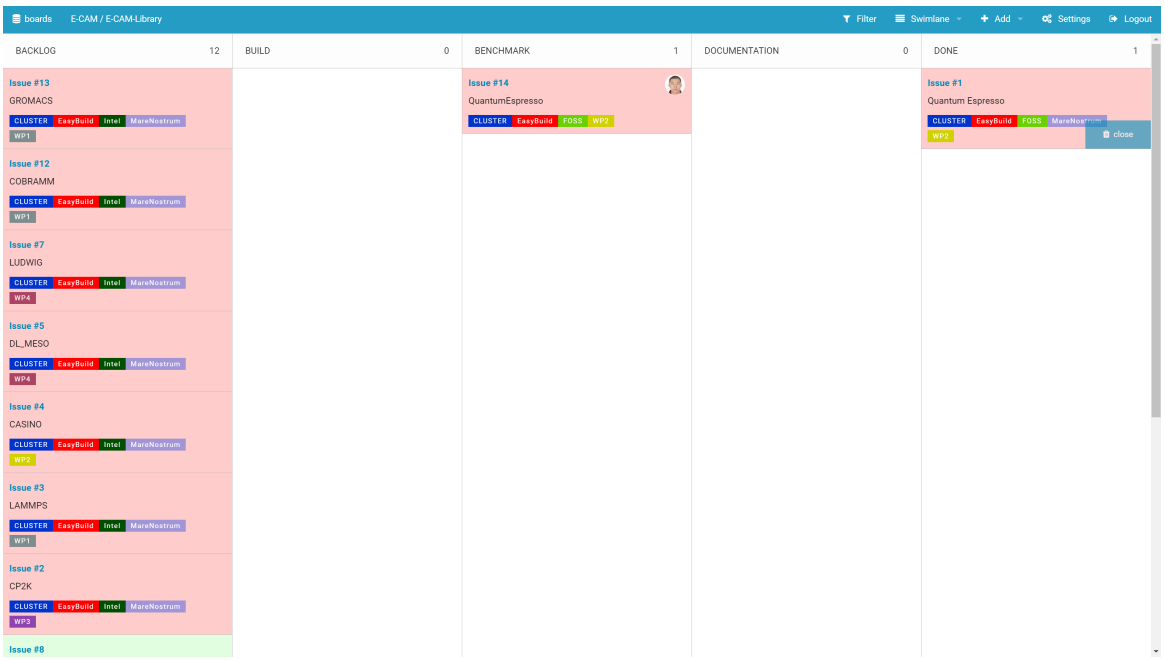

Figure 4: An example use case for our Kanban service

3.3.3 EtherPad

In geographically distributed collaborative projects, regular meetings are essential. To facilitate taking minutes in such meetings we have created an Etherpad service. Etherpad is a web-based collaborative real-time editor, allowing authors to simultaneously edit a text document, and see all of the participants' edits in real-time, with the ability to display each author's text in their own color.

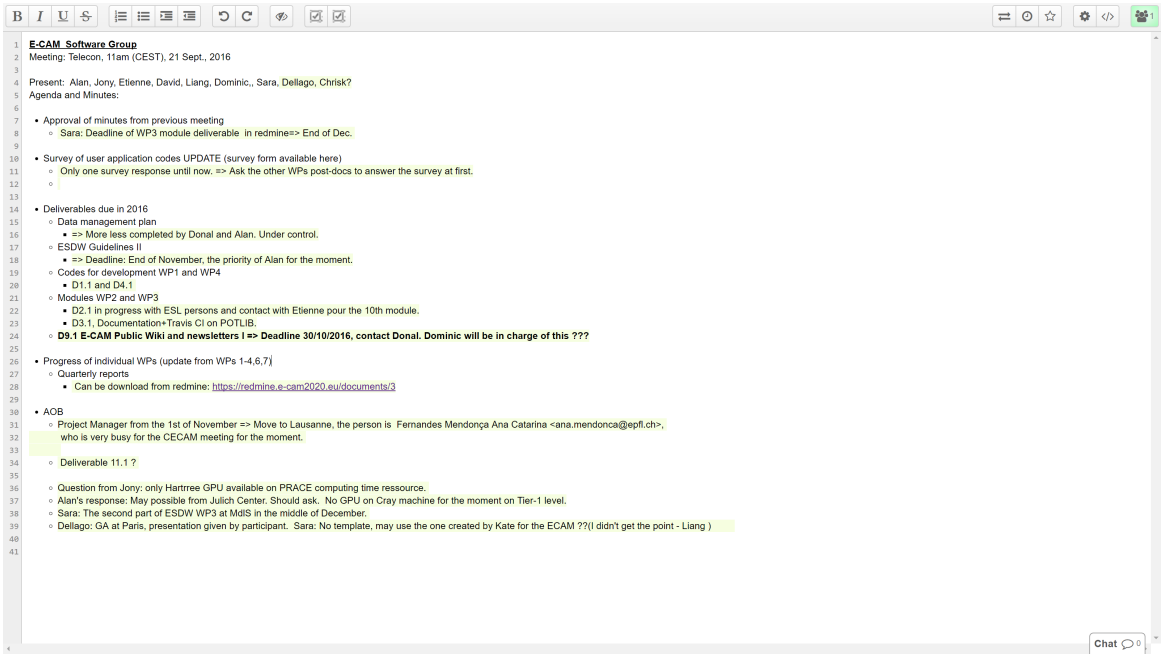

Figure 5: An example use case for our Etherpad service

An example use of an etherpad is given in Fig. 5.

3.3.4 ShareLatex

Many of the E-CAM beneficiaries are academic researchers, who use Latex to produce technical documents and papers. ShareLaTeX is an online LaTeX editor that allows real-time collaboration and online compiling of projects to PDF

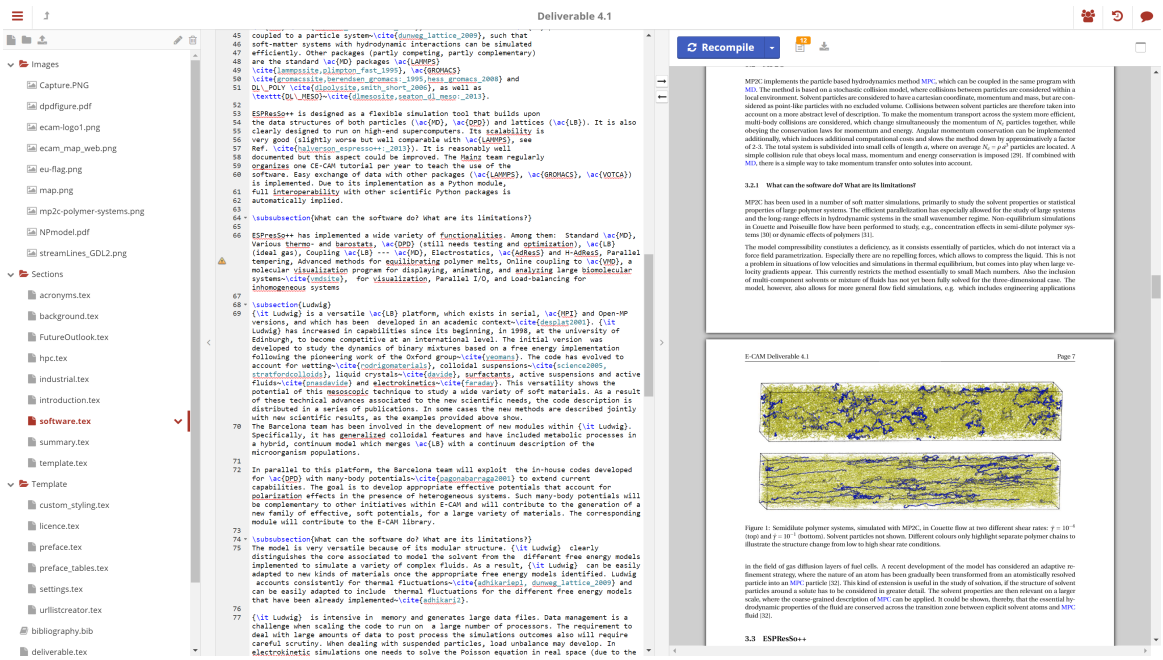

Figure 6: An example use case for our ShareLatex service

format. In comparison to other LaTeX editors, ShareLaTeX is a server-based application, which is accessed through a web browser. We have created a ShareLatex service for use by the E-CAM community to facilitate easier collaboration on project-related papers. This tool is also used in the production of all E-CAM deliverables. An example can be seen in Fig. 6

4 Web Infrastructure for Teaching Tools

E-CAM wishes to develop an appropriate online training infrastructure over the course of the project. To this end, it is establishing strong partnerships with PRACE and other similar initiatives in Europe and the US in order to provide innovative training content that can bring the E-CAM user communities to extreme scale.

The infrastructure to support these efforts is now in place with the initial content collected during the ESDW program of 2017 currently being added. The extent of the development efforts to create the platform, as well as the organisational structure we adopt to leverage it, is outlined in the following subsections.

The results of this effort, the E-CAM Training Portal, are publicly available since December 2017.

4.1 Lecture Capture

For lecture capture, video gamers have fantastic (and fantastically cheap) hardware and software options: in our case we use the Live Gamer Portable capture card for screen capture and the open source Open Broadcast Software (which is capable of processing multiple sources, combining them and live streaming them result). We have created a fully portable capture kit to capture training events made up of just a laptop, the Live Gamer Portable to capture the presenters screen, an HD webcam and a wireless microphone for speaker and audience.

4.1.1 Remote participation

The final combined video stream can also be passed as a screen share via, e.g., a Google Hangout (for public lectures) or a service such as GoToMeeting (for limited private participation). With this approach the remote participant can engage in realtime interaction with the speaker and/or local audience.

We also use the E-CAM Etherpad service to take live notes of the event. This tool has a chat window where any participant can ask questions without interrupting the flow of the speaker (which can be very useful if they may have, for example, missed a terminal command on a previous slide).

We intend to advertise remote participation to our ESDW events in the 2018 program, and also to encourage workshop organizers to promote it among their contacts.

4.1.2 Example of captured content

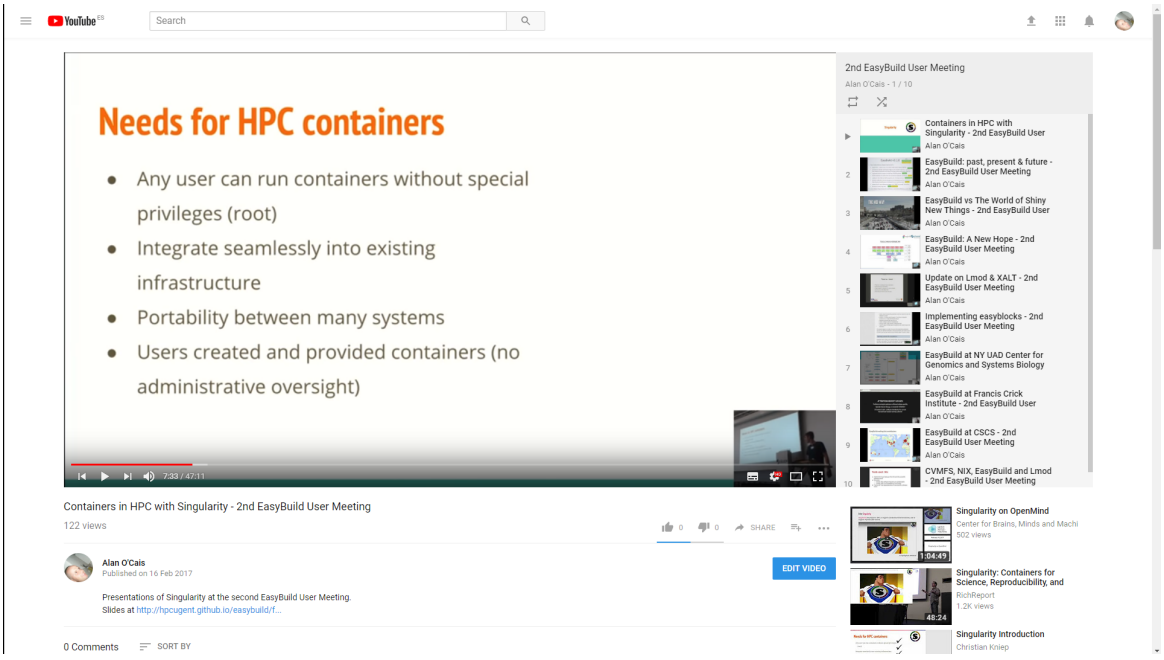

Figure 7: A captured lecture and YouTube playlist of an entire event.

We conducted a trial run of all of these facilities during the 2nd EasyBuild User Meeting that was co-organised by E-CAM.

We successfully captured all lectures with only minor technical issues with respect to MAC compatibility with the capture card. We even had a remote lecture with the lecturer able to discuss directly with the audience, the only technical issue was poor quality of the internet connection at one point (which was outside of our control). The full program of the event with links to the captured content is available at the 2nd EasyBuild User Meeting website.

The captured lectures are published on YouTube and the playlist can be seen in Figure 7 (which includes an example of what a captured lecture looks like).

4.1.3 Capture of E-CAM events

High quality content that will translate well to our online training portal has been captured at all our ESDW events in 2017. After an initial evaluation of the effort involved and the measured impact, we expect to extend this to all our events in 2018.

4.2 Content Management

It is most definitely not purely the content generated by E-CAM which is of interest to the E-CAM community. In addition to managing our own content, we also want to be able to include material from other sources (whether it be video content, online material or PDF documents). We want to store and treat this content (video, document or URL) in a consistent manner. We also want the ability to associate additional material with each content type (source code for tutorials is an obvious example).

The content that is generated by E-CAM is captured with the permission of the speakers, in many cases the speakers are not comfortable with the material being publicly available but are willing to make the content available to the E-CAM user community. This ability to restrict who can access material places a further requirement on the selection of a content management system. Once the video has been captured we would like the ability to extract additional information about the video, in particular we would like to automatically create navigation for the presentation video (since many presentations are one hour or longer).

These requirements have brought us to Clowder, which is developed at National Centre for Supercomputing Applications (NCSA). Clowder is a research data management system designed to support any data format and multiple research domains. Clowder contains three major extension points: preprocessing, processing and previewing. When new data is added to the system, whether it is via the web front-end, or through the RESTful web services, preprocessing is off-loaded to extraction services for extracting appropriate data and metadata. The extraction services attempt to extract information and run preprocessing steps based on the type of the data, for example to create previews. This raw metadata is presented to the user in the Clowder web interface.

Clowder was not originally intended to be leveraged as a pedagogical infrastructure, E-CAM chose to use it this way because creating and maintaining, for example, a set of Massive Online Open Course (MOOC)s is simply too labour intensive and E-CAM does not have the resources for this. The automation possibilities that Clowder presents allow us to drastically reduce the end-user overhead for adding content (or at the very least bring it to a level that E-CAM can afford to sustain). Of course the end result is not as pretty as it's labour intensive alternatives but it provides us with a low-effort dynamic service not easily found with other approaches.

The ability to add new file types and develop our own extraction services that act on them to create new metadata are seen as very valuable to our infrastructure. In the short term it has allowed us to create a computer vision algorithm to extract and store slide transitions alongside video material, but we also foresee further use cases (for example, automated parsing of EasyBuild configuration files to create URL-based cross-referencing, syntax highlighting of uploaded source code,...). The automation aspect of the extraction services ensure that E-CAM should be able to expend a controlled amount of effort on adding new data even when the amount of new data is large.

The logical structure of Clowder is also of benefit to E-CAM. In section 4.3 we go into greater detail about how we embed our concept into the Clowder infrastructure. We call the E-CAM clowder instance the E-CAM Training Portal.

4.2.1 Outsourcing

Much of the development work for our Clowder extractor services has been outsourced. An experienced computer scientist was required to understand and implement the requirements that E-CAM has outlined, and his work was completed in December 2017. Nevertheless, the Clowder infrastructure is still under active development, mainly by the E-CAM software manager.

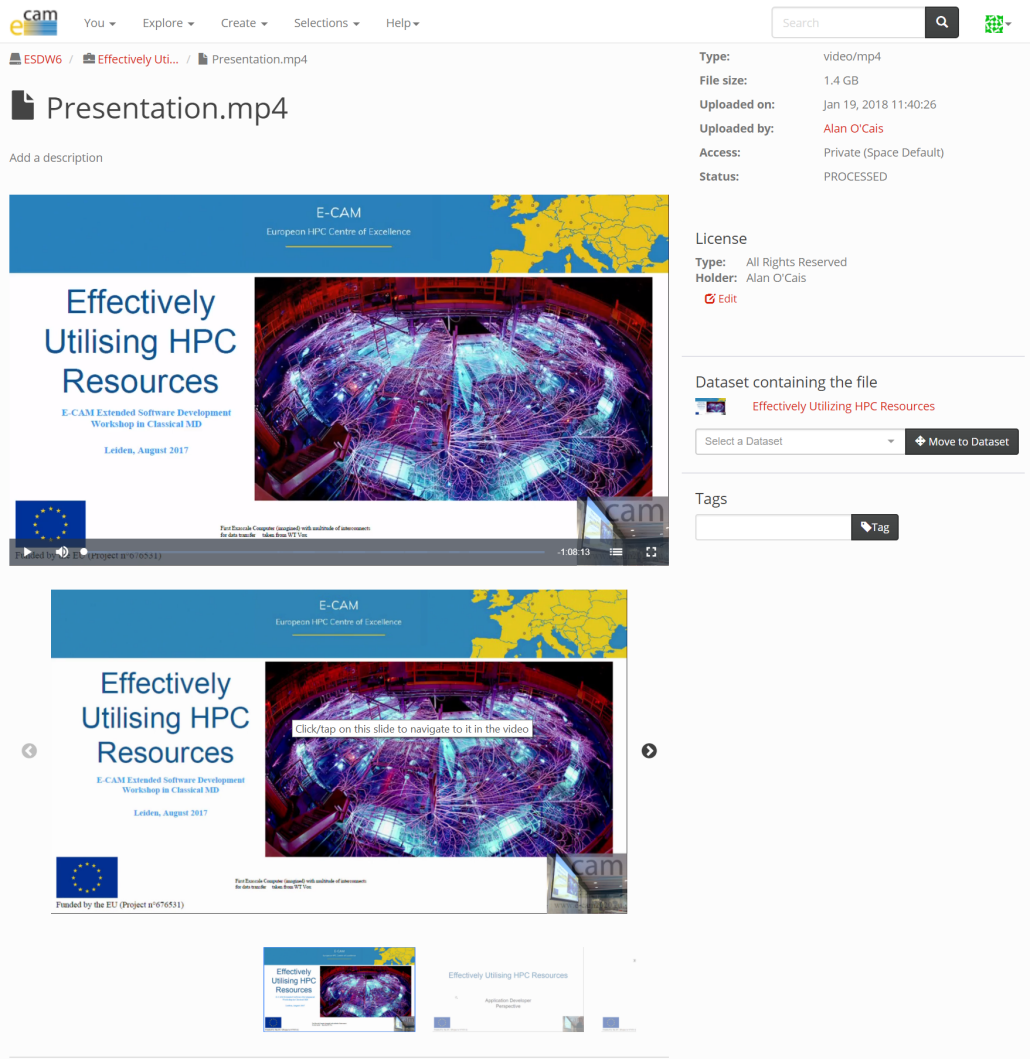

Figure 8: Captured presentation with navigation carousel.

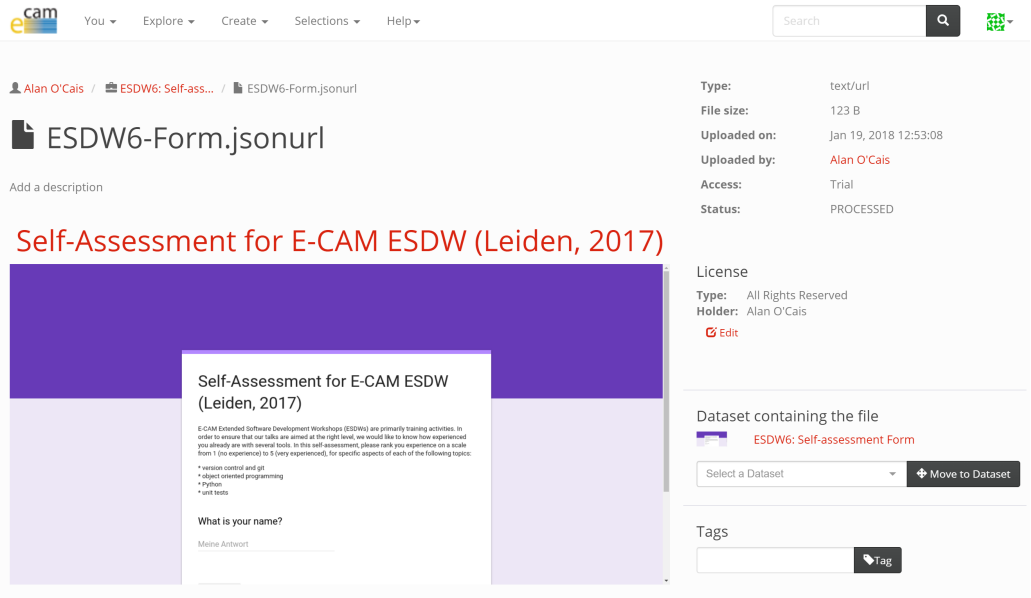

Figure 9: Example of a URL in Clowder, in this case the self-assessment form for an ESDW.

All code generated from these development is open source and can be found at:

- Video presentation extractor and previewer - This has the capabilities to extract slide timestamps and images from a captured presentation and prepare them for the previewer. It also shrinks the input video to just 1.2MB

per minute for full HD video (roughly the same size as simple stereo audio files). This is possible with extremely heavy compression suitable for the particular makeup of a presentation. The previewer creates a navigation panel for the video allowing the user to easily jump between slides in the video. An example of an extracted presentation and its display in the previewer is shown in Fig. 8

- URL extractor and previewer - This extractor grabs a screenshot of the target URL (as well as some additional site information) and the previewer displays this within Clowder. Upon clicking on the image if possible the target URL is opened in a modal window[9] in the same page, if not the URL is opened in a new tab. An example a URL in the previewer can be seen in Fig. 9 (in this case the URL is a link to the pre-assessment questionnaire for an ESDW)

4.3 Structuring the Training Portal

The logical structure of Clowder can be easily leveraged by E-CAM. Single *files* are collected into *datasets*, which are a group of files that through some defined relationship or corresponding metadata are strongly tied together and not representable otherwise by the individual files. *Datasets* are what we expect our captured lectures to be (made up of the associated videos, lecture materials, reading materials, tutorial content and software requirements). *Collections* are a user defined group of *datasets* and other *collections*. A *space* is a group of *collections*, *datasets*, and *files* with defined user access rights. Spaces are used to share data within datasets and collections with other users, in our case this will be how we will restrict access to (some of) our material.

A schema of how we will organize the information in Clowder is represented in Fig. 10. Below we go into greater detail about how we embed our concept into the Clowder infrastructure.

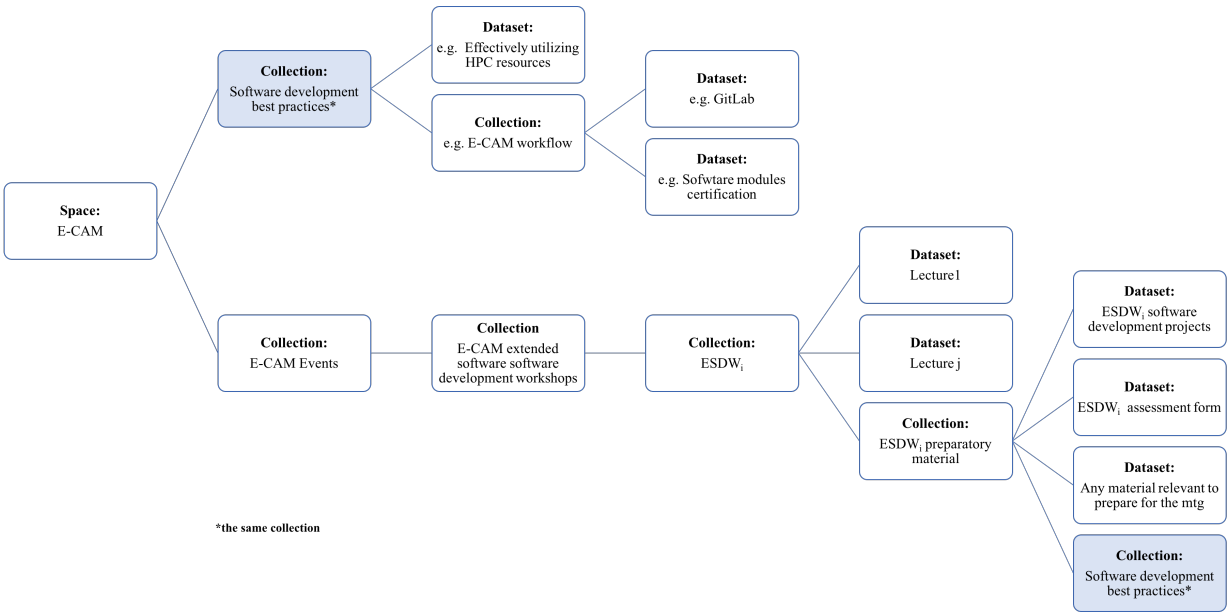

Figure 10: Schema of how we will organise the training material generated by E-CAM in the Clowder infrastructure.

4.3.1 Datasets

The site has a simple interface to upload data. The metadata extracted by Clowder depends on the file extension of the file uploaded, the current files that are "understood" by Clowder are mp4/webm, jsonurl (as described below) and pdf. Additional filetypes can be added but this may require development work (since Clowder needs to be told what type of metadata you want to extract for the new filetype and how to extract it). A file is uploaded as (part of) a *dataset*.

The basic concepts of an **extractor** and **previewer** are relevant to files, and are explained as follows:

- Once a file with a known filetype is uploaded, Clowder passes it to an extractor service associated with the filetype. This service extracts metadata related to the file.

[9] A modal window is a popup window that is displayed on top of the active webpage

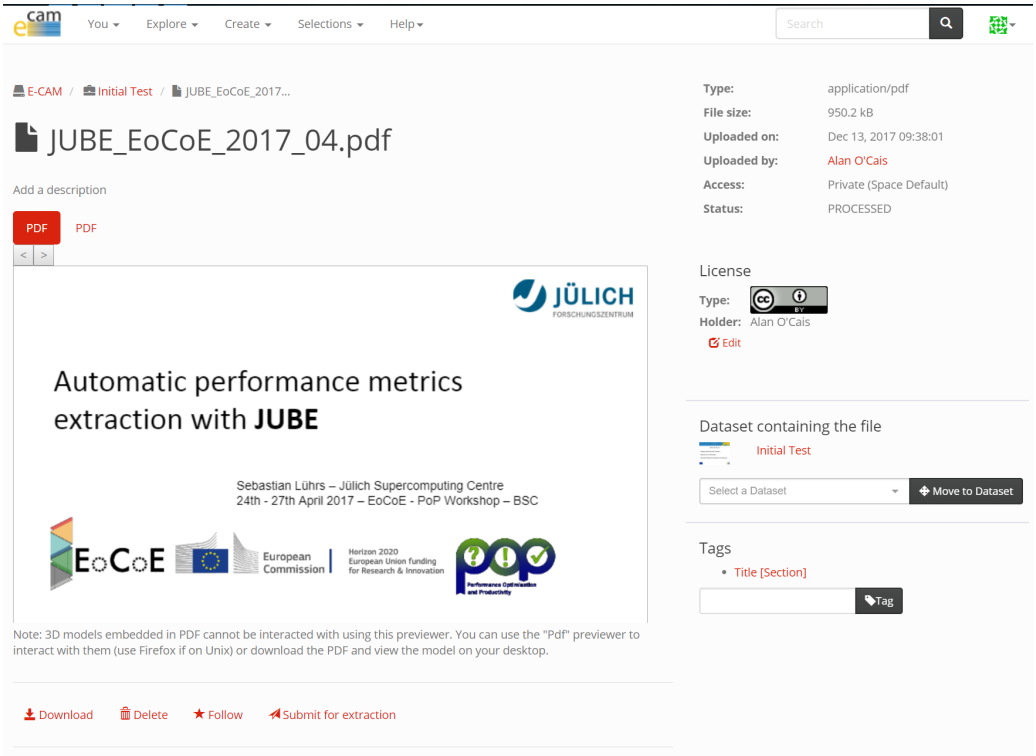

Figure 11: A PDF file rendered within Clowder.

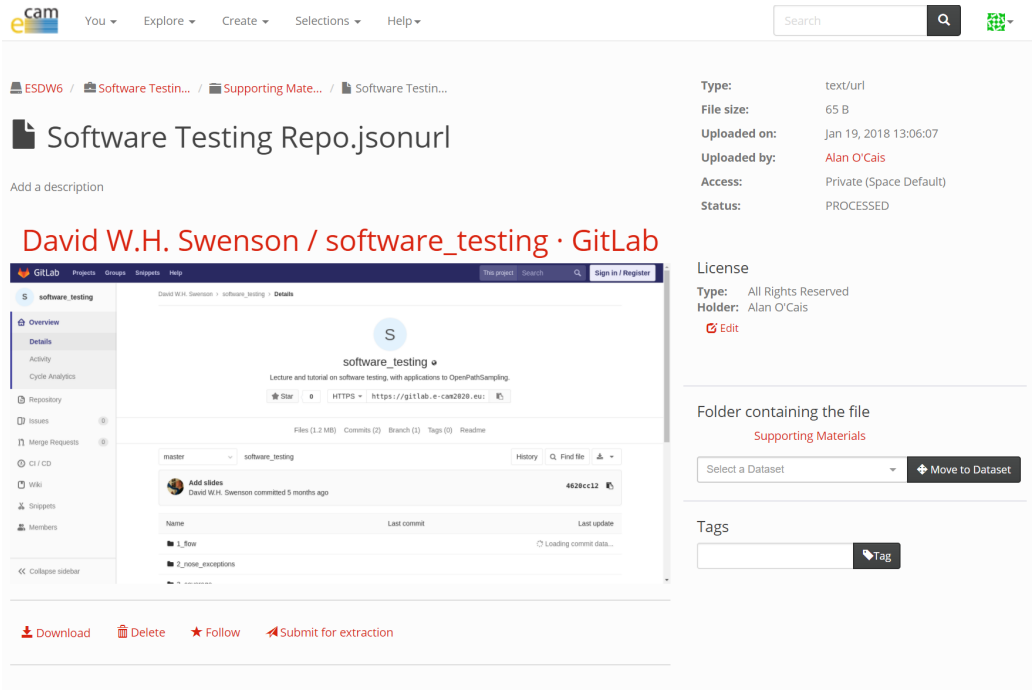

Figure 12: A repo provided as a link to GitLab.

– For a pdf, the metadata is simply a thumbnail image of the front page (but one could consider, for example, running character recognition over the entire document and uploading the result as additional metadata).

– For a jsonurl we figure out whether the url can be served over https and whether it can be "framed". We also take a screenshot of the website to use as a thumbnail. For a URL, you need to provide a URL in a JSON format via a plaintext file with the file extension 'jsonurl'. An example of such a "file" represented in Clowder can be seen in Fig. 12. The contents of the file should look like:

```
{
  "URL": "https://gitlab.e-cam2020.eu/e-cam/E-CAM-Library"
}
```

```
5
```

We have also created a simply utility to help you create this file. Currently this tool is located on W3Schools until we migrate it to a permanent home.

  - The extractor for a presentation video is the most complex. We use computer vision methods to identify slide transitions in the presentation (typically created as described in Section 4.1). The time of each transition is registered and an image is captured to represent the slide. We also reencode the video with very heavy compression so that it can be easily viewed through the website. This extractor can take a long time to gather the metadata (up to 3 times the length of the video).

- Once the metadata has been created, Clowder can leverage it in a previewer to visually represent the data in the website.
  - For a pdf, the previewer shows the document itself within the webpage. The previewer allows you to mark a particular area of a page of the pdf, display it seperately and make a direct comment/description. For an example see Fig. 11.
  - For a URL, a screenshot of the relevant website is displayed. This screenshot can be clicked and if it is "frameable" it will be opened within the same page in a new frame. If it is not "frameable", clicking the link will open it in a new tab. For an example see Fig. 12.
  - For a video, the highly compressed version is displayed with the slides underneath. The slide can be used to navigate through the presentation. With the video in full screen mode, there is a special tab in the bottom right that can be used to navigate through the slides. For an example see Fig. 8.

Of particular note, one should consider that the potentials for leveraging URLs are limitless: quizzes, youtube, articles, webpages, google docs,dropbox files,...You should be able to leverage anything available on the web with this approach (and this is the obvious approach to link to software repositories, another example of which can be seen in Fig. 12).

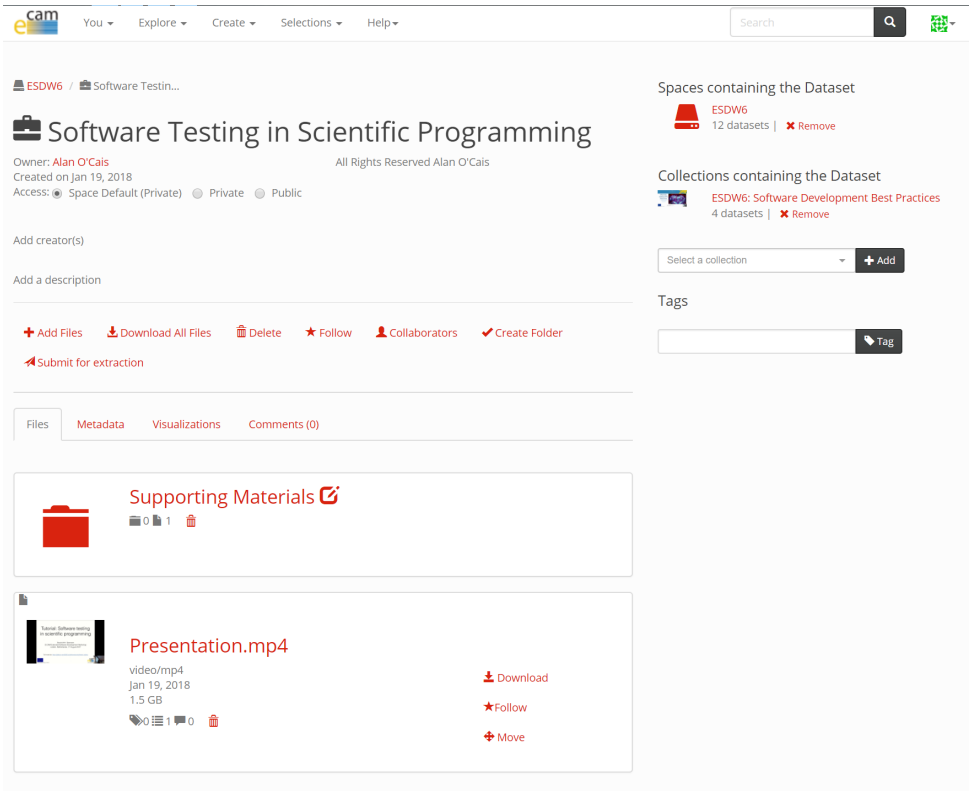

Figure 13: An example of lecture dataset which includes a directory for supporting material (in this case the Git repository of Fig. 12).

With respect to IP, a rights holder can be defined for each file. It is also possible to associate a licence with each file. Existing licence types are *Limited*, *Creative Commons* and *Public Domain*. With a limited licence, one can define a custom licence (and provide a weblink to a detailed description). This licence also allows the possibility of disallowing downloading of the raw file (but the extracted metadata will still be viewable). It is possible to select variants of

the Creative Commons licence based on your particular requirements. With a Public Domain licence, you are not restricting how the data can be reused.

Files in Clowder are part of a dataset. It is typical to have one file in a dataset but this is not a hard restriction. For example you may present a lecture divided in two parts, with slides available as pdfs and a URL link to a software repo where the practical material can be found. A dataset should basically form a logical unit (but what that strictly means is something we will have to gradually define as we gain practical experience).

It is also possible to create folders within datasets to help organise the files within a dataset. An example of such organisation of a dataset can be seen in Fig. 13.

There are limited ways to alter how files in a dataset are listed: by age, alphabetical order or their reverse. When a lecture is uploaded in parts, it is recommended to add "Part 1", "Part 2", etc. to the title to facilitate sorting.

4.3.2 Collections

Datasets can be gathered together into collections. An example use might be to collect together lecture datasets that made up a workshop, as can be seen in Fig. 14 where we gather the set of lecture datasets from an ESDW together into a collection. In reality we go further than this and create a *Child Collection* from the lectures presented that were purely to do with software development, as seen in Fig. 15. This distinction is helpful for the user as it allows them to easily focus on either the scientific or software development aspects of the ESDW.

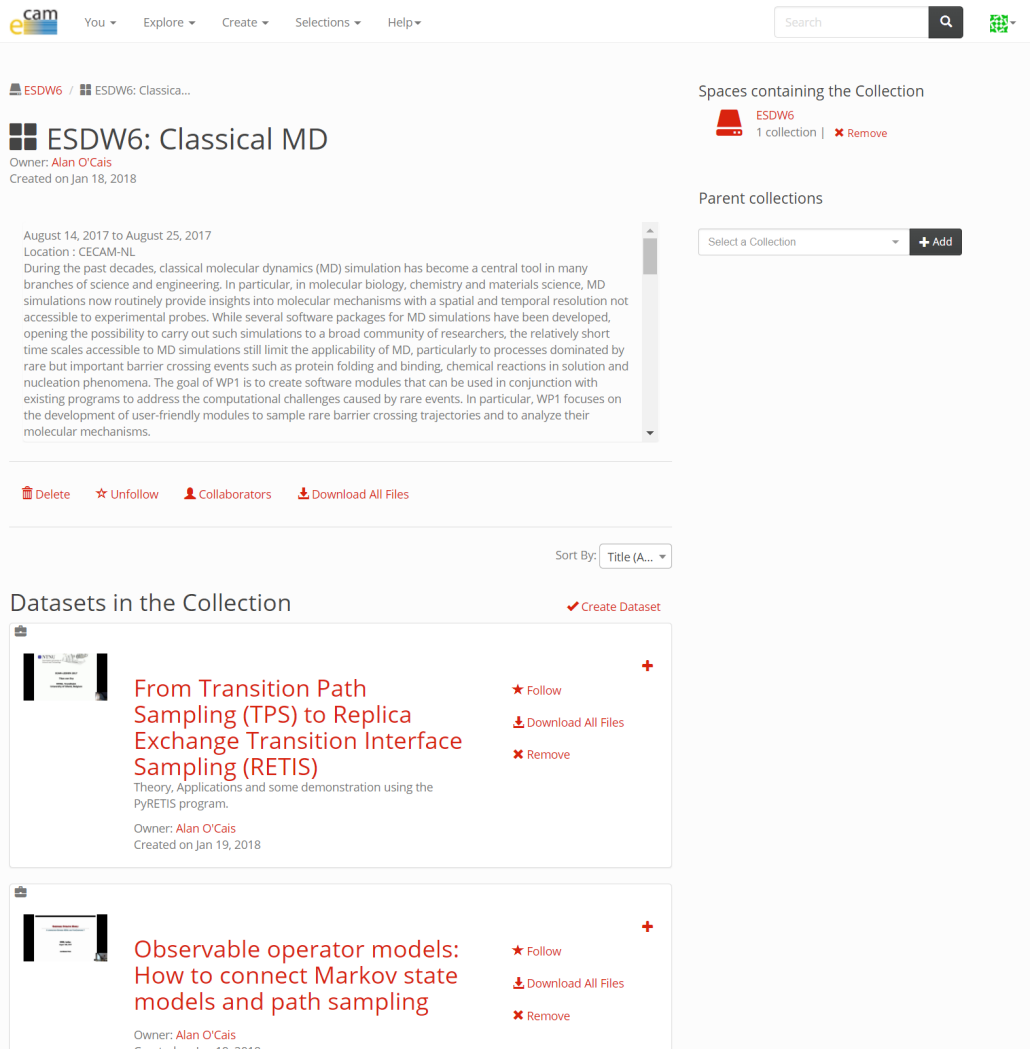

Figure 14: A set of lectures are gathered together into a collection.

Furthermore you can have collections of collections. So, for example, one can imagine creating a collection of all the ESDWs in a given year. There are no restrictions on the number of collections a dataset/collection can appear in so the possibilities are vast. Our immediate use case is to create a core set of "Software Development Best Practices" that

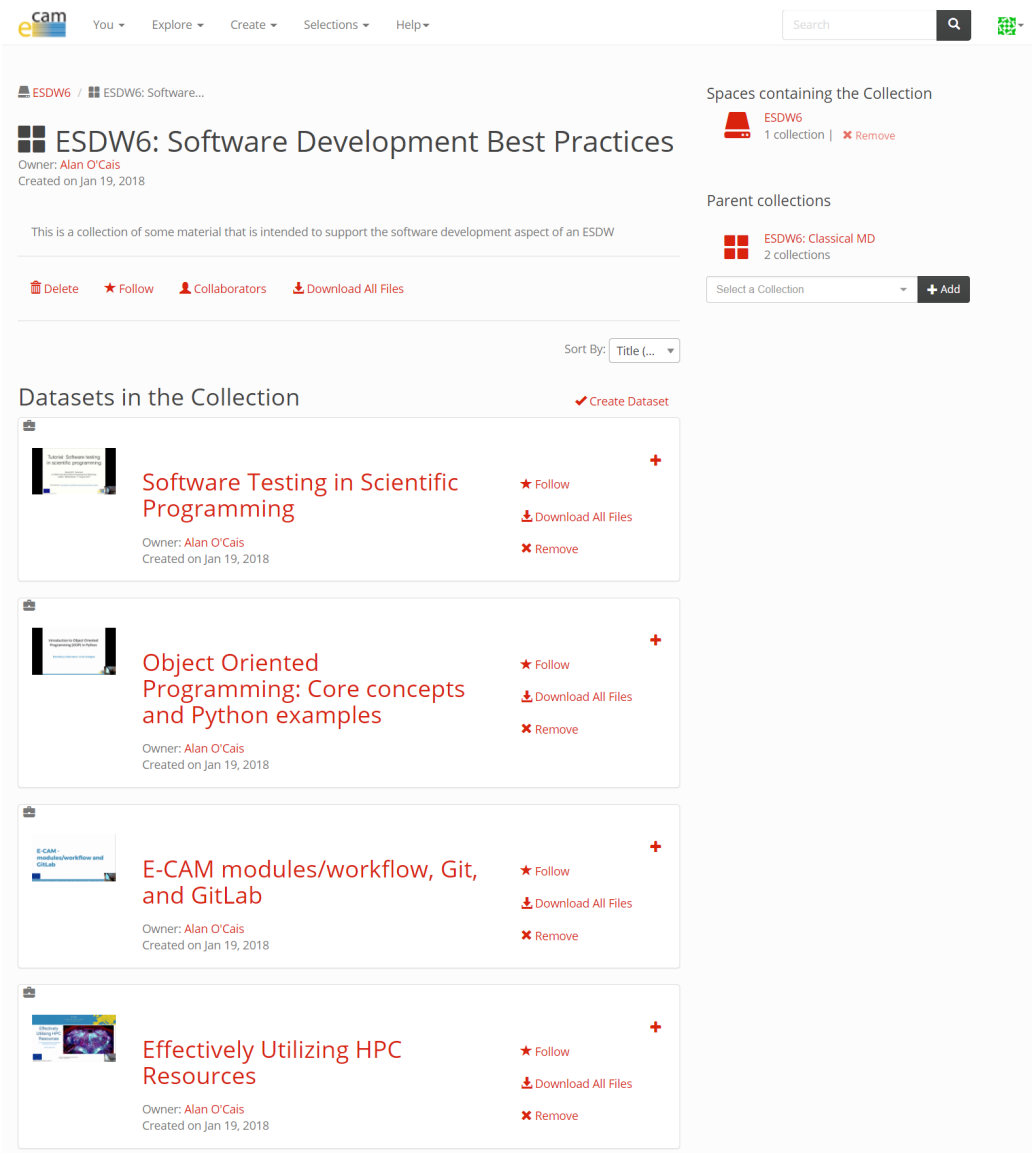

Figure 15: A child collection created from the set of lectures relating to software development within the ESDW.

can be recycled for each ESDW. This allows us to create new material at each ESDW rather than have to repeatedly deliver the same material to a new audience.

4.3.3 Spaces

Finally, you have collaboration *spaces* within Clowder. A *spaces* is a group of *collections*, *datasets*, and files with defined user access rights. In other words, *datasets* and *collections* can sit in a space which is used to collaborate with a particular group. An example of a set of spaces can be seen in Fig. 16.

You can configure whether you want your *datasets/collections/spaces* to be public or private. If a private *dataset* is part of a *spaces* , people with access to the *spaces* can view the file. Any user can request access to a *spaces* , but this request must be approved by a space administrator. **Public files can be found directly via weblink (no login required) even if they are in a private space.**

These facilities allow us to share within the E-CAM community all our content in addition to providing a subset that is completely public. The privacy feature allows us the possibility of placing content under embargo (as sometimes people may present information that they have not yet fully exploited).

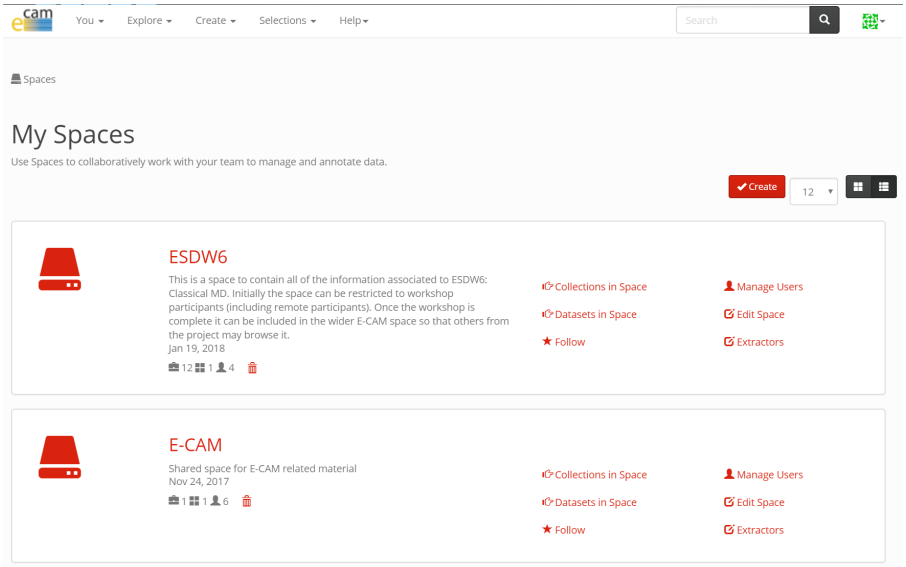

Figure 16: An example of set of spaces. The "ESDW6" space is for attendees to the associated ESDW, its contents can be transferred to the "E-CAM" space (which has many more members) once the workshop is fully completed.

4.4 Future Development

Future efforts will largely focus on expanding the availability of content within the portal and leveraging it effectively during ESDW events. The final development possibility we are considering is:

- High Performance Computing (HPC) services - this service would provide an HPC environment to train users and to test modules before moving the testing onto larger PRACE machines. There is a possibility to create a pseudo-HPC-cluster using Docker containers and provide access to this over the internet.

This effort would have significant development overhead as no such drop-in service currently exists, and we are carefully weighing the benefit of this service against the cost of provision.

In 2018, E-CAM will focus on streaming portions of the training events that it is associated to. The goal is to allow remote participation for E-CAM members, and in particular for industrial partners who cannot commit to the allocated time for our ESDW events.

4.4.1 Collaboration with other projects

E-CAM has committed to collaborating with PRACE on some of the training courses in it's PRACE Advanced Training Centre courses. In particular, we collaborate with JSC in 2018 on an upcoming "High-performance scientific computing in C++" and "Python in High Performance Computing" courses.

Furthermore we are contact with the MolSSI which is a US-based project that aims to serve as a nexus for science, education, and cooperation serving the worldwide community of computational molecular scientists – a broad field including of biomolecular simulation, quantum chemistry, and materials science. We are investigating the possibility of collaboration from a training perspective given the clear overlap of the scope of our projects.

References

Acronyms Used

CECAM Centre Européen de Calcul Atomique et Moléculaire

HPC High Performance Computing

PRACE Partnership for Advanced Computing in Europe

ESDW Extended Software Development Workshop

WP Work Package

IP Intellectual Property

PDRA Postdoctoral Research Associate

PATC PRACE Advanced Training Centre

MolSSI Molecular Sciences Software Institute

MOOC Massive Online Open Course

IP Intellectual Property

NCSA National Centre for Supercomputing Applications

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/37
a.ocais@fz-juelich.de ... mailto:a.ocais@fz-juelich.de
http://creativecommons.org/licenses/by/4.0 ... http://creativecommons.org/licenses/by/4.0

Page iii

E-CAM Software Library ... http://e-cam.readthedocs.io/en/latest/
GitLab service ... https://gitlab.e-cam2020.eu/
Redmine service ... https://redmine.e-cam2020.eu/
Kanban service ... https://kanban.e-cam2020.eu/
Etherpad service ... https://etherpad.e-cam2020.eu/
ShareLatex service ... https://sharelatex.e-cam2020.eu/
captured lecture and YouTube playlist of an entire event ... https://www.youtube.com/playlist?list=PLVA9BuLC1j-yfxp2w-wraAGDCmhjb3o5Y

Page 1

E-CAM project website ... https://www.e-cam2020.eu/
software modules of E-CAM ... https://www.e-cam2020.eu/software-library/
upcoming E-CAM events ... https://www.e-cam2020.eu/events
E-CAM online services ... https://www.e-cam2020.eu/resources/
technical first contact page ... https://www.e-cam2020.eu/technical-questions/
Type C PRACE Preparatory Access ... http://www.prace-ri.eu/prace-preparatory-access/
detailed descriptions of the E-CAM events ... https://www.e-cam2020.eu/events/list/?tribe_paged=1&tribe_event_display=past
E-CAM Training Portal ... https://clowder.e-cam2020.eu
Clowder ... https://clowder.ncsa.illinois.edu/
NCSA ... http://www.ncsa.illinois.edu/
E-CAM GitLab service ... https://gitlab.e-cam2020.eu/
ReadTheDocs.org ... https://readthedocs.org/
Kanban service ... https://kanban.e-cam2020.eu/
Redmine service ... https://redmine.e-cam2020.eu/
Etherpad service ... https://etherpad.e-cam2020.eu/
ShareLatex service ... https://sharelatex.e-cam2020.eu/

Page 3

E-CAM Software Platform II . . . https://www.e-cam2020.eu/wp-content/uploads/2017/10/D6.4_12102017.pdf
E-CAM Software Platform II . . . https://www.e-cam2020.eu/wp-content/uploads/2017/10/D6.4_12102017.pdf
Clowder authentication plugin for GitLab . . . https://github.com/ncsa/clowder/pull/3
E-CAM project website . . . https://www.e-cam2020.eu/

Page 4

E-CAM Software Library . . . http://e-cam.readthedocs.io/en/latest/
E-CAM Software Library . . . http://e-cam.readthedocs.io/en/latest/
https://gitlab.e-cam2020.eu/e-cam/E-CAM-Library . . . https://gitlab.e-cam2020.eu/e-cam/E-CAM-Library
WP1 Deprecated Site . . . http://e-cam-classical-md-modules.readthedocs.io/en/latest/index.html
WP1 Current Webpage . . . http://e-cam.readthedocs.io/en/latest/Classical-MD-Modules/index.html#readme-classical-md
WP2 Deprecated Site . . . http://e-cam-electronic-structure-modules.readthedocs.io/en/latest/
WP2 Current Webpage . . . http://e-cam.readthedocs.io/en/latest/Electronic-Structure-Modules/index.html#readme-electronic-structure
WP3 Deprecated Site . . . http://e-cam-quantum-dynamics-modules.readthedocs.io/en/latest/
WP3 Current Webpage . . . http://e-cam.readthedocs.io/en/latest/Quantum-Dynamics-Modules/index.html#readme-quantum-dynamics
WP4 Deprecated Site . . . http://e-cam-meso-and-multi-scale-modelling-modules.readthedocs.io/en/latest/
WP4 Current Webpage . . . http://e-cam.readthedocs.io/en/latest/Meso-Multi-Scale-Modelling-Modules/index.html#readme-meso-multi

Page 5

Sphinx documentation generator . . . http://www.sphinx-doc.org/en/stable/index.html
E-CAM GitLab service . . . https://gitlab.e-cam2020.eu/e-cam/E-CAM-Library
ReadTheDocs.org . . . https://readthedocs.org/
Module Contribution Guidelines . . . http://e-cam.readthedocs.io/en/latest/contributing.html

Page 6

E-CAM project website . . . https://www.e-cam2020.eu/
E-CAM events calendar . . . https://www.e-cam2020.eu/events
E-CAM online services . . . https://www.e-cam2020.eu/resources/
Submitted deliverables . . . https://www.e-cam2020.eu/deliverables/
E-CAM publications . . . https://www.e-cam2020.eu/publications/
Workshops scientific reports . . . https://www.e-cam2020.eu/scientific-reports/
Software repositories . . . https://www.e-cam2020.eu/software-library/
Pilot projects with industry . . . https://www.e-cam2020.eu/pilot-projects-with-industry/
Job opportunities . . . https://www.e-cam2020.eu/vacancies/
Latest news items . . . https://www.e-cam2020.eu/category/news/
technical first contact page . . . https://www.e-cam2020.eu/technical-questions/
Type C PRACE Preparatory Access . . . http://www.prace-ri.eu/prace-preparatory-access/
detailed descriptions of the E-CAM events . . . https://www.e-cam2020.eu/?post_type=tribe_events
webpage . . . https://www.e-cam2020.eu/pilot-projects-with-industry/
E-CAM GitLab service . . . https://gitlab.e-cam2020.eu/

Page 7

D6.2: E-CAM software development tools . . . https://www.e-cam2020.eu/wp-content/uploads/2017/09/D6.2_Resubmit_04092017.pdf
GitLab service . . . https://gitlab.e-cam2020.eu/
all publicly accessible repositories of E-CAM . . . http://gitlab.e-cam2020.eu/explore/projects
clone of E-CAM documentation pages . . . http://e-cam.e-cam2020.io/E-CAM-Library

Page 8

Redmine . . . http://www.redmine.org/
Redmine service . . . https://redmine.e-cam2020.eu/
Redmine service . . . https://redmine.e-cam2020.eu/
Kanban . . . http://kanbanblog.com/explained/

Page 9

Kanban service . . . https://kanban.e-cam2020.eu/

Etherpad service ... https://etherpad.e-cam2020.eu/
Etherpad service ... https://etherpad.e-cam2020.eu/

Page 10

ShareLatex service ... https://sharelatex.e-cam2020.eu/
ShareLatex service ... https://sharelatex.e-cam2020.eu/

Page 11

E-CAM Training Portal ... https://clowder.e-cam2020.eu
Live Gamer Portable ... http://gamerzone.avermedia.com/game_capture/live_gamer_portable
Open Broadcast Software ... https://obsproject.com/
Live Gamer Portable ... http://gamerzone.avermedia.com/game_capture/live_gamer_portable
GoToMeeting ... https://www.gotomeeting.com/
E-CAM Etherpad service ... https://etherpad.e-cam2020.eu/
captured lecture and YouTube playlist of an entire event ... https://www.youtube.com/playlist?list=PLVA9BuLC1j-yfxp2w-wraAGDCmhjb3o5Y
2nd EasyBuild User Meeting ... https://github.com/easybuilders/easybuild/wiki/2nd-EasyBuild-User-Meeting

Page 12

2nd EasyBuild User Meeting website ... https://github.com/easybuilders/easybuild/wiki/2nd-EasyBuild-User-M
captured lectures are published on YouTube ... https://www.youtube.com/playlist?list=PLVA9BuLC1j-yfxp2w-wraA
Clowder ... https://clowder.ncsa.illinois.edu/
NCSA ... http://www.ncsa.illinois.edu/
E-CAM Training Portal ... https://clowder.e-cam2020.eu

Page 13

Video presentation extractor and previewer ... https://github.com/E-CAM/training_portal

Page 14

URL extractor and previewer ... https://github.com/E-CAM/url_extractor

Page 16

W3Schools ... https://www.w3schools.com/code/tryit.asp?filename=FN4KT9YMQUHT

Citations

[1] A. O. Cais, "E-CAM Software Platform I," Nov. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.841751