D1.2: Data-management Plan

INTERTWinE Work Package 1

Project Acronym	INTERTWinE
Project Title	Programming Model INTERoperability ToWards Exascale
Project Number	671602
Instrument	Horizon 2020
Thematic Priority	FETHPC

Due date	End of PM6 (31st March 2016)
Submission date	30/MAR/2016
Project start date	01/OCT/2015
Project duration	36 months
Deliverable lead organization	UEDIN
Version	1.4
Status	Published
Author(s)	George Beckett (UEDIN)
Reviewer(s)	Olivier Aumage (Inria); Daniel Holmes (UEDIN); Toni Collis (UEDIN)

Dissemination level	
PU	*Public*

Copyright INTERTWinE Consortium 2018

Version History

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	26/FEB/16	Skeleton plan created using template on Digital Curation Centre website	MGB
0.2	14/MAR/16	Plan populated for each data type	MGB
0.3	24/MAR/16	Revisions to source-code management, in response to reviewers comment	DH, MGB
1.0	30/MAR/16	Version uploaded to EC	MGB
1.1	21/APR/16	Additional information as to the management of BSCW service provide.	MGB
1.2	15/AUG/16	Added information about transmission of personal data to outside of EU	MGB
1.3	29/MAR/17	Revision to handling of demographic data, in response to requirements of WHPC	MGB
1.4	29/MAY/18	Updated to reflect process changes introduced to meet GDPR requirements	MGB

Copyright INTERTWinE Consortium 2018

Table of Contents

Copyright INTERTWinE Consortium 2018

1 Introduction

This deliverable contains the data-management plan for the INTERTWinE project. The plan has been developed using the template for Horizon 2020 projects that is published on the Digital Curation Centre website. The plan is reviewed and, if necessary, updated at key points in the project—at least, once every eighteen months.

The plan documents the approach taken to manage each significant dataset that is the responsibility of the project (see Section 2) as well as providing guidance to project team members on relevant data-management matters (see Section 3).

1.1 Purpose

The data-management plan identifies the key forms of data that the project team will create and manage. For each of these data forms, guidance is provided on how the data should be formatted/ marked-up and curated, aiming to ensure that:

- It is curated with sufficient provenance to support typical reuse (for example, reproduction of experiments and follow-up enquiries).
- It is easily reusable by others both within and outside of the project (as appropriate).
- It is managed in a secure manner, based on the value and sensitivity of the data.

Further, the expected use of each data form is documented, along with guidelines as to applications, media, and tools that are well-suited to the usage that is anticipated.

All members of the INTERTWinE project team should be familiar with this (and subsequent versions of the) data-management plan, and should adhere to the guidelines in their project-related work.

1.2 Glossary of Acronyms

API Application Programming Interface

CSV Comma-separated Values

PDF (Adobe's) Portable Document Format

Copyright INTERTWinE Consortium 2018

2 Data-management Plan

2.1 Administrative Details

Project Name: INTERTWinE

Project Identifier: INTERTWINE-H2020

Grant Title: 671602

Principal Investigator / Researcher: Mark Bull (University of Edinburgh)

Description: The INTERTWinE project addresses the problem of programming-model design and implementation for the Exascale. One of the main challenges is in the interoperability of application programming interfaces (APIs). This project seeks to address this interoperability, bringing together the principal European organisations driving the evolution of programming models and their implementations. In the course of the project, the team will create research reports, software source code, and training materials that document and disseminate their findings. The team will also operate a training program, for which they will collate registration data and training feedback, to assist them to better understand the market for the training and to determine improvements to the training material.

Funder: European Commission (Horizon 2020)

Institutions: University of Edinburgh; BSC; Inria; KTH; Fraunhofer; DLR; T-Systems SFR; Universitat Jaume I de Castellon; University of Manchester

2.2 Data Collection – Slides for Training Courses

Data set reference and name
Slides for Training Courses

Data set description
Training material, in the form of PowerPoint slide decks, will be created by members of the project team to convey key supercomputing programming skills and techniques to computational scientists. Each deck may incorporate material from other training presentations, subject to meeting licensing restrictions and attribution requirements.

Standards and metadata
Slide decks will be created in Microsoft PowerPoint (pptx) format, using the INTERTWinE project presentation template. Slides will be generated following topical guidelines to ensure equality of access for all to training. Further, slides will be limited to content that can be exported to Adobe Portable Document Format (PDF) without loss of meaning—for example, avoiding use of animations to overlay new content on top of older content in a slide transition.

Data sharing
Slide decks will be owned by the Consortium (accepting any third-party material that is used), in line with the terms of the Consortium Agreement. Slide decks will be presented by INTERTWinE staff members, or approved third-parties, at face-to-face training events. Slide decks may also be converted to PDF format for distribution to training course attendees (via email or project website) and the public (project website), with a suitable licence.

Archiving and preservation (including storage and backup)
Slide decks will be held on the project's BSCW server until at least one year after the planned end of the project (at the time of writing, this is end of September 2019). It is anticipated that each slide deck will occupy 1—20 Megabytes of disk space and that the project will produce no more than 20 such slide decks. The project holds a licence

Copyright INTERTWinE Consortium 2018

for BSCW [1] that runs for one year longer than the planned project duration. The project's BSCW server is backed-up to an Edinburgh University Information Services-managed backup system at a separate location and on a nightly basis

2.3 Data Collection – Feedback from Training Courses (including participant demographic information)

Data set reference and name

Feedback from Training Courses

Data set description

As part of the registration process and at the end of each training course, we will circulate a feedback form to course participants. Completion of this form is optional. The feedback form will be hosted online, using a GDPR-compliant survey tool. Participants will complete the form via the survey-tool's web interface.

Shortly after each training course, the feedback will be downloaded from the survey tool as a CSV-format table. The feedback will then be deleted from the survey tool using the service's administrative controls.

As the feedback form contains personal information (participant's demographic information—specifically gender and age), the data is subject to GPDR and needs to be managed accordingly. A privacy statement is to be provided along with the feedback form and the privacy statement must fulfil GDPR requirements. The feedback forms will be stored in an access-controlled section of the project's BSCW server, with access limited to those specific project team members that are responsible for analysing the demographic data.

The dataset must not be distributed beyond the project unless specific restrictions are fulfilled: either the data must be completely anonymised to remove personal information; or a contract must be in place between the INTERTWinE lead organisation and the third-party describing what can and cannot be done with the data, and in line with GDPR.

Standards and metadata

Both course feedback (original and anonymised versions) and aggregated demographic data will be stored in CSV files. The course instance (course name, venue, and date) that a particular CSV file relates to will be recorded in the tail of the CSV file.

Data sharing

It is expected that, to fulfil INTERTWinE's ambitions to promote diversity and redress the gender imbalance in the HPC community, feedback from training courses will be shared with Women In HPC [4] and similar organisations, to be used only to help with understanding the demographics of who is attending training courses in relation to their experiences. Careful consideration of GDPR must be made before sharing data, and it may be necessary to engage a GDPR expert to review and approve the proposed approach. The data will never be used to identify individuals and will only be reported in an aggregate, anonymised format.

Other feedback (not containing personally identifiable information) will be shared with course organisers and course presenters, contributing to the improvement process for training courses. Feedback, in anonymised summary form, may also be used on publicity material and project reports to demonstrate the quality and popularity of the courses. Feedback must be anonymised as it contains personal information.

The feedback that is collected will include personal information about EU citizens. Transmission of this personal information is regulated by European law and, in particular, as the survey-tool provider may store the feedback on servers outside of the

Copyright INTERTWinE Consortium 2018

EU, we need to make everyone who provides feedback aware of this situation in advance. To do this, every feedback form will include a checkbox—near the top of the form—advising of the potential transmission of their responses to a server outside of the EU. Further, the form should not allow feedback to be submitted unless this box is checked.

Archiving and preservation (including storage and backup)

The CSV files holding course feedback and aggregated demographic data will be held in an access-controlled area on the project's BSCW server, with access limited to the Work Package 2 team (and, by their nature, the service administrator). The CSV files are estimated to require no more than 10 Megabytes to store. The project holds a licence for BSCW that runs for one year longer than the planned project duration. The project's BSCW server is backed-up to an Edinburgh University Information Services-managed backup system on a nightly basis

2.4 Data Collection – Project Materials

Data set reference and name

Project Materials

Data set description

During the course of the project, various documents will be prepared for internal use—including, for example, meeting minutes, briefing notes, team-member contact information, and process documents.

This material will be held in the project's document repository (BSCW), which implements fine-grain access control and version control. The repository is organised as a tree-like hierarchy of folders organised by work package (Work Packages 1—5) or by governance entity (for example, the Executive Group and Board).

Standards and metadata

The Project Manual [1] details the standard formats that should be used for creating and distributing different kinds of document. BSCW's metadata is used to track authorship and contributions, last modified time and versioning.

Data sharing

By default, all Project Materials will be treated as project internal and will be protected by access-control mechanism built into BSCW. For those materials that are to be shared more widely (notably final versions of slide presentation, white papers, and dissemination material), the Project Manual details in what form these material may be distributed (usually PDF) and in line with the terms of the Consortium Agreement [2].

Project Material that contains personally-identifiable information should not be distributed outside of the project team, unless there is a genuine business interest for so doing (e.g. noting authors of a document and their institutional affiliation). Project Material containing personally identifiable information should typically be shared as a link to the copy held in BSCW rather than as an email attachment or via another file-sharing service.

Archiving and preservation (including storage and backup)

The project holds a licence for BSCW that runs for one year longer than the planned project duration. A space quota of 30 Gigabytes has been configured for the project on a University of Edinburgh-managed storage system. The project's BSCW server is backed-up to a second, independent system on a nightly basis, and then replicated to the University of Edinburgh-managed tape backup system (with a 60-day retention policy).

Copyright INTERTWinE Consortium 2018

The various servers involved in offering the BSCW service are patched, with fixes and updates from the O/S vendor, on a weekly basis. Further, the BSCW service is updated as and when updates are issued by OrbiTeam (the BSCW vendor).

2.5 Data Collection – Benchmarking Results

Data set reference and name

Benchmarking Results

Data set description

Measuring the performance of software developed by the project, and of reference applications that use this software, is a key device for the project to assess the effectiveness of its output. Benchmark measurements (usually run-time or scalability efficiency) will be recorded for each significant revision to project software, and will be curated in an agreed form for easy comparison and cross-reference.

For all benchmarks, the native output (usually text files in a format determined by the benchmark author), will be captured and stored. Output files containing raw data will also be created and stored for each modifiable benchmark. In addition, any programs or scripts used for analysis of these raw data files, and needed for reproducibility (plus the resulting processed data files) will be stored.

Native output files, raw data files, processed data files, and the source code for analysis programs will be stored in the project's source-code repository—along with the source code for the benchmark, if available.

As much as possible, the benchmarking operation will be automated—using scripts to record the precise details of the system environment, to build the required software, to execute the chosen benchmark programs, to extract the resulting measurements, and to perform analysis using these results. Benchmarking will be manually triggered by an operator when a significant revision to project software has been created. These scripts will be treated as any other project software, following the process detailed in the Project Manual [1].

Standards and metadata

At the time of each benchmark run, the operator will use a script to take careful note of the experiment parameters (including context information—such as names of those involved, date and time—plus specific information about the versions/ source/ build parameters of any software used in the experiment, the hardware/ platforms used, and any options or switches passed to the benchmark process). This metadata will be recorded in a suitable text file and added to the source-code repository at the same time as the significant revision to the project software—to maintain the association.

Data sharing

We envisage that selected benchmarking results will be used widely to demonstrate project outputs and impact, as well as being included in resulting publications. The project team may choose to share benchmarking tools with other groups (outside of the project team) to allow project results to be checked or comparative investigations to be undertaken, according to the terms of the Consortium Agreement [2].

Archiving and preservation (including storage and backup)

Benchmark results will be stored in the project's source-code repository. The University of Edinburgh hosts the project's GitLab repository and will continue to do so for at least one year longer than the planned project duration. The project's GitLab repository is backed-up to an Edinburgh University Information Services-managed backup system, at a different location and on a nightly basis.

Copyright INTERTWinE Consortium 2018

2.6 Data Collection – Project News Mailing List

Data set reference and name

INTERTWinE News mailing list

Data set description

The project maintains a mailing list of email addresses (and, in some cases, names) of people who are interested to hear occasional (typically, monthly or less frequent) updates on project progress or closely related topics from third-parties.

The mailing list is held on the University of Edinburgh's Mlist service, which: is based on the Sympa technology; holds subscriber details on a secure server; provides a password-protected web interface for administering the list and its membership; provides an email interface for posting news items, and for individuals to adjust/ delete their subscription information.

The mailing list is administered by the Project Manager and the Work Package 2 Leader.

The list is operated in a self-sign-up mode. Individuals must sign themselves up to the list: the project team do not add email addresses on behalf of others. Individuals may unsubscribe themselves from the list at any time, or ask one of the mailing-list administrators to unsubscribe them.

Standards and metadata

The format of the subscriber list and the email archive is the default for the Sympa tool.

Data sharing

The subscriber list contains personally identifiable information (email addresses and potentially names) and is therefore subject to GDPR. A privacy statement should be accessible—included (or linked) in each message posted to the list and published on the website. The subscriber list should not be downloaded from Mlist nor shared with people outside of the INTERTWinE project team.

Posts to the mailing list may only be made by one of the mailing-list administrators and should not contain significant personal information, except where there is a legitimate business interest (e.g. names of training presenters, author lists for papers). The mailing list is designated as a private list on Mlist, which is only visible to subscribers and designated administrators from the project team.

Archiving and preservation (including storage and backup)

An archive of postings to the list is maintained for three years. This is accessible to those active subscribers who have an Mlist web-account, via the Mlist web interface. In practice, this means members of the University of Edinburgh who are subscribers or mailing-list administrators.

The mailing list should be deleted within 12 months of the INTERTWinE project ending (that is, before 30th September 2019).

Copyright INTERTWinE Consortium 2018

3 Other Considerations

3.1 Transmission of EU citizens' personal data outside of the EU

The European Commission prohibits the transfer of personal information about EU citizens outside of the EU, unless the destination has adequate security in place to protect privacy. To make it easier for European companies to use US-based services (for example, cloud services), a Safe Harbour agreement was enacted between the USA and EU in 2000, which allowed US companies to self-certify their compliance to the EU's definition of adequate privacy protection. However, in October 2015, the European Court of Justice ruled that the Safe Harbour agreement was invalid, necessitating EU companies to make other arrangements in order to legitimately transmit data about EU citizens to the US. It is important to be aware of these regulations when engaging third-party services to support project activities.

Colleagues from the EUDAT project [https://www.eudat.eu/] have advised that it is sufficient to ensure that anyone providing personal information that may be transmitted outside of the EU is aware of this in advance of providing the information.

At the time of writing, two services have been checked with regard to transmission of personal information:

- SurveyMonkey – which is used to collect feedback from training-course participants. A legitimate mechanism for collecting feedback using SurveyMonkey is descripted in Section 2.3.
- GotoMeeting – which is used to run distributed project meetings and which includes a 'chat client' functionality that could involve the transmission of personal information to outside of the EU. At the time of writing, it has been confirmed that the chat-client functionality is a peer-to-peer protocol only involving the clients that are connected to a particular meeting: nothing from the chat client is transmitted to or stored on GotoMeeting servers.

US and EC authorities are, at the time of writing, in discussions regarding a replacement for the Safe Harbour Agreement. The project team will monitor the progress of these discussions and update this plan if and when appropriate.

3.2 Impact of GDPR

In addition to the items considered in Section 2, the project team should always be vigilant to the possibility that activities will be subject to GDPR. In particular, the following activities should be raised with the Project Manager, in advance:

- **Creating a new project mailing list**—since this will involve collecting and managing personal information (email addresses and possibly names);
- **Circulating surveys and web forms**—since this may involve the collection, handling and potentially transmission of personally identifiable information, as well as the use of third-party services.
- **Collection of website analytics data**—Google Analytics, which is used to monitor traffic to the project website, can involve the collection and transmission of personally identifiable information (e.g. specific location information about visitors; IP addresses of web clients). The following must be done, in relation to Google Analytics:
    - Transmission of personally identifiable data to Google Analytics should be disabled.
    - IP anonymisation should be enabled in Google Analytics.
    - The use of cookies should be avoided and, if using cookies, a Privacy Policy should be provided.

Copyright INTERTWinE Consortium 2018

4 References

[1] George Beckett, *INTERTWinE Project Manual*, https://intertwine-bscw.epcc.ed.ac.uk/sec/bscw.cgi/d7045-3/*/*/*/intertwine_project_manual.html (project-internal, please contact george.beckett@ed.ac.uk for further information).

[2] INTERTWinE Consortium, *INTERTWinE Consortium Agreement*, 15th December 2015

[3] OrbiTeam Software GmbH & Co. KG, *BSCW – Basic Support for Collaborative Working*, http://www.bscw.de/ (website accessed on 23rd March 2016).

[4] SurveyMonkey, *SurveyMonkey*, http://www.surveymonkey.com (website accessed on 23rd March 2016).

[5] Women in HPC, *WHPC*, http://www.womeninhpc.org.uk/ (website accessed on 23rd March 2016).

Copyright INTERTWinE Consortium 2018