Directorate General for Communications Networks, Content and Technology

Innovation Action

ICT-687655

D1.3– Data Management Plan

Due date of deliverable: 30 November 2017

Date delivered: 17 January 2018

Start date of project: 1 December 2015 Duration: 36 months

Lead contractor for this deliverable:

Version: 2.0

Confidentiality status: **Public**

Abstract

For a multidisciplinary and multicultural collaborative project like 2-IMMERSE, quality assurance of its deliverables and publications is very important. This deliverable describes the high-level quality assurance measures that the 2-IMMERSE project intends to apply. The details of the quality measures of the technical components and the 2-IMMERSE software are outside the scope of this deliverable and described in the particular technical deliverables.

Besides 2-IMMERSE internal quality assurance mechanisms there are regular and irregular reviews and audits performed by the European Commission. These are planned and controlled by the Commission, and are not the main subject of this deliverable. However, Annex B gives an overview of the Commission performed reviews and audits.

This version provides a 'mid term' check on whether the cited procedures are being adhered to, and finds that they are.

Target audience

2-IMMERSE team members, the Project Officer from the European Commission, 2-IMMERSE reviewers and auditors

Disclaimer

This document contains material, which is the copyright of certain 2-IMMERSE consortium parties, and may not be reproduced or copied without permission. All 2-IMMERSE consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the 2-IMMERSE consortium as a whole, nor a certain party of the 2-IMMERSE consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

Impressum

Full project title: 2-IMMERSE

Title of the workpackage:

Document title:

Editor : Peter Stansfield, BBC

Workpackage Leader: Mark Lomas, BBC

Technical Project Leader: Mark Lomas, BBC

Project Co-ordinator: Helene Waters, BBC

This project is co-funded by the European Union through the Horizon 2020 programme.

Executive Summary

This deliverable describes all of the data generated and collected within the 2-IMMERSE project. For the dataset descriptions we used the official guidelines on Data Management Plan (DMP) from the Horizon 2020 portal. This document is the first iteration and cannot be considered as a final document. It will evolve and gain more focus as the project progresses and we learn from our experiments and pilots.

The contents of this deliverable will inform the work of the work packages specifically WP3 and WP4 where experimental data will be collected and published using these guidelines. 2-IMMERSE will produce a number of technical results resulting from the deployment of 4 pilots and the Dissemination Plan D6.1 discusses the means of promoting those results to the research community and across the industry. The main elements of that plan are open access to scientific publications and open-source releases of 2-Immerse platform software.

2-IMMERSE is participating in the Horizon 2020 Pilot on Open Research Data to make its research data available. This document describes data to be shared, associated metadata and how the data will be stored and made available.

This version, updated on 9th January 2018 provides a measure of how the ‘will do’ actions have been turned into ‘have done’ actions. Whilst not all of them would be expected to show this transition, it is reassuring that several have.

List of Authors

Phil Stenton, Peter Stansfield BBC

Reviewers

Hélène Waters, BBC

Table of Contents

1 Data Summary

Data will be collected during experiments and pilots (field trails) exploring the provision and value of multi-screen experiences of drama and sport in private and public venues. WP3 will deal with experiments to inform the design of technology between trials and the address the general user questions that straddle the 4 pilots and WP4 will deal with the design of the technology and user experience for each pilot. An example WP3 experiment is the viability of video chat technology in the situations we envisage with multiple sound sources across multiple locations on the size and power of platforms we are likely to deploy across the 4 trials. WP4 would examine the value and delivery of things like a scrolling script on a second screen during the broadcast of a Shakespeare play.

Sone data has now been collected, and the table shows which of the recommended storage and management plans has been adopted where appropriate.

1.1 Purpose

The data will be collected to inform experience design and technology development within and across the pilots: from setting up and configuring multi-screen environments, through signing up for services (on-boarding), enjoying the service and closing down. Guidelines that can be extrapolated beyond the pilot scenarios will also be noted and made public.

Data may also include code in the form of composeable micro-services called Distributed Media Applications clusters.

1.2 Reuse

The data generated by each pilot and its preceding experiments will be evaluated to assess the success of the assumptions and designs supported by the implementation of the technology and the delivery of the experience. The insights drawn will be published along with summary data. Anonymised raw data may help other researchers draw further conclusions through re-analysis, comparison or combination with other data.

1.3 Origin of the Data

The data will be derived from subjects in lab experiments and workshops and the four pilots. Data will be in the form of quantitative measurements through data analytics, likert scales and video logging over the course of the trials and qualitative data via feedback questionnaires and structured interviews and/or focus groups before and after the event (e.g. Shakespeare play, football match). Data will be provided by members of the public taking part in trials in homes and public venues.

Code as data will be developed over the course of the project both to support the trials and the interactions within them and the general architecture across trials to support flexible and customisable multi-screen environments.

1.4 Expected Size and Utility

The data will be in the form of transcribed interviews and questionnaire responses, data analytics of screen usage and device interactions and video recording of video chat and observation. This behavioural data may be useful to anyone wishing to understand the potential for multiscreen environments to enhance the experience of drama and sport. In addition researchers or broadcast practitioners may be interested in considering the data to learn general lessons across other genres of entertainment.

2 Fair Data

Quantitative and qualitative data (for example, ratings and transcribed comments) obtained from participants in user research within 2-IMMERSE, will be stored safely and retained for no longer than is necessary for the purposes of the research and in accordance with the goals of the EU Data Pilot. This will form part of the participation agreement to which participants must give informed consent prior to involvement. Any data which could identify an individual participant (including name, age or demographic class) will be encrypted, stored separately from evaluation data, and retained for no longer than is necessary for the purposes of the research.

The related consent form will include the fact that a video is being made and will be looked at in future stages of the research. Video files will be encrypted before any transfer away from the lab where they were made, and not made available to any organisation outside the consortium, which will also be on the consent form. All data will be handled in accordance with both EU regulations around data protection, and national government regulation in the country where the study takes place.

2.1 Making data findable, including provisions for metadata

Transcripts of interviews will be created where possible using Kaldi http://kaldi-asr.org/ a speech to text system used by the BBC in its Snippets retrieval system and for the recovery of subtitles.

Computer Assisted Qualitative Data Analysis Software CAQDAS will be used to manage this data. Both qualitative and quantitative data will be made available through publications and through metadata tagged on-line datasets where possible.

Code specifications will be released through project deliverables. The 2-IMMERSE system architecture is built from a number of defined services, each scoped with specific roles and responsibilities. These services will be designed to scale elastically, running the required number of instances to meet dynamic load requirements. The array of services types and instances need to collaborate and interoperate with each other to deliver the 2-IMMERSE experience. As described in D2.1 the project was originally going to use Mantl, http://docs.mantl.io/en/latest/ a modern platform for rapidly deploying globally distributed services. Mantl provides an integrated set of industry-standard open-source components. It is cloud infrastructure provider agnostic, and can be deployed on AWS, OpenStack, Vagrant, Bare Metal etc. Mantl is licensed by 2-Immerse partner Cisco under the Apache Version 2 License. However we have now switched to Rancher, which we feel more suitable to our needs. See https://rancher.com/rancher/

2.2 Making data openly accessible

The question “*Which data produced and/or used in the project will be made openly available as the default?”,* will be outlined later in the project as the research unfolds. Our default intention is to make the results of the project available through Open Source repositories. We are exploring the use of OpenAIRE https://www.openaire.eu/ repository https://zenodo.org/ Zenodo. However, in practice we found it easier to clean the code manually, document it, and put it into Github.

2.3 Making data interoperable

Our aim is to make the data we produce in the project interoperable, allowing data exchange

and re-use between researchers and institutions. How big a task this is will become clearer as the architecture requirements emerge. One of the goals of the project is to provide a functioning implemention of HbbTV 2.0 and inform the next generation of that standard. Data management of the code created will be the responsibility of the technical team under Cisco's leadership and the user data management will be the responsibility of the social science team lead by the BBC. Resources for long term preservation of the user data have not yet been discussed. Much will depend on the perceived significance of the data and the provisions under the Data Protection Act 1998 https://www.gov.uk/data-protection/the-data-protection-act.

2.4 Data security

Data when generated will be kept in secure encrypted repositories within the partners' organisations. When it is released for wider availability it will be through secure stores such as OpenAIRE on the CERN sponsored Zenodo facility.

2.5 Ethical aspects

Data will be anonymised so that subjects in the trials cannot be identified through their participation and informed consent forms will include the permission requests for the long term storage and public sharing of their data. Sensitive data may include conversations during a home or school theatre event or home or public venue sporting events. If Video, audio and transcription data is made available for sharing it will be anonymised before doing so. Video may not be shared if permission is refused or if identities cannot be confidently hidden.

3 Progress in the second year

We summarise here in tabular form the promises and achievements in the second year of the project

Section	Promise	Did we do it?
1.0	Data will be collected during experiments and pilots (field trails) exploring the provision and value of multi-screen experiences of drama and sport in private and public venues	First Theatre at Home & MotoGP trials data has been collected from pilots. In addition data from a Lab test exploring young people's responses to different presentations of Football have also been collected.
1.2	The data generated by each pilot and its preceding experiments will be evaluated to assess the success of the assumptions and designs supported by the implementation of the technology and the delivery of the experience.	We have completed analysis of the Theatre at Home data and of the Young peoples' perceptions of different presentations of football on TV. We have some data for the MotoGP at Home evaluation but this is still being collected (as of Jan 2018) and has not yet been analysed. Analysis for the MotoGP data should take place in February 2018.
1.2	The insights drawn will be published along with summary data. Anonymised raw data may help other researchers draw further conclusions through re-analysis, comparison or combination with other data.	Project deliverables and industry/academic conference papers referenced data summaries and raw data. All data was anonymised in these documents.
1.4	The data will be in the form of transcribed interviews and questionnaire responses, data analytics of screen usage and device interactions and video recording of video chat and observation	Theatre at Home used interview data (stored in password protected files on an external hardrive, locked in a cupboard in BBC offices); online survey data (collected by survey monkey & password protected); and behaviour data (password protected files). All data is anonymised –using participant IDs not real names. Young peoples' perceptions of different presentations of football on TV has been written up for an NEM conference and presented. The data and work will be submitted in a project deliverable related to WP3 in due course – when we discuss in more detail the design thinking work that went into the Football prototype service MotoGP trials –survey & interview data is stored in password protected files. All data is anonymised – using participant IDs. There are no video recordings of observations yet.

Section	Promise	Did we do it?
		But we may video record sessions in the BBC/BT user testing labs –which will be stored in password protected files.
1.4	This behavioural data may be useful to anyone wishing to understand the potential for multiscreen environments to enhance the experience of drama and sport. In addition researchers or broadcast practitioners may be interested in considering the data to learn general lessons across other genres of entertainment.	To date –none of the behavioural data from Theatre at Home or MotoGP has been shared with anyone outside the project team –to re-analyse. Summaries of behaviour data has been shared via industry/academic conference papers. MotoGP data may be shared with BT Sport and BBC Sport in the future.
2.0	Quantitative and qualitative data (for example, ratings and transcribed comments) obtained from participants in user research within 2-IMMERSE, will be stored safely and retained for no longer than is necessary for the purposes of the research and in accordance with the goals of the EU Data Pilot.	All data is anonymised, and password protected. During the trials it is stored on BSCW and locally on a project hard drive, (stored in a metal secured cupboard in BBC offices). Once the 2Immerse project is complete all data will be transferred to OpenAIRE (and deleted from hard drive storage).
2.0	Any data which could identify an individual participant (including name, age or demographic class) will be encrypted, stored separately from evaluation data, and retained for no longer than is necessary for the purposes of the research.	For Theatre at Home & MotoGP data which could identify participants is kept separately from evaluation data. Files are encrypted. These files are deleted once the analysis has been completed, as they are no longer useful.
2.0	All data will be handled in accordance with both EU regulations around data protection, and national government regulation in the country where the study takes place.	We are checking the most recent EU regulations on data protection, but believe we are compliant.
2.1	Transcripts of interviews will be created using Kaldi http://kaldi-asr.org/ a speech to text system used by the BBC in its Snippets retrieval system and for the recovery of subtitles.	For Theatre in Home Kaldi wasn't accurate enough for all participants, so we took audio recordings and transcribed the data manually. Audio files and transcriptions are stored on BSCW. No identifying features were used –we used first names and participants IDs. (No surnames, addresses or phone

Section	Promise	Did we do it?
		numbers, etc were recorded.) For MotoGP –interviewers from Acumen transcribe the data during the interviews. No audio recordings have been taken. Participants who may complete the trials in BT/BBC will be video recorded but participant IDs and/or first names will be used, and files will be password protected. For the assessment of different presentations of football presentations by young people this work did not involve interviews, only scores.
2.1	Computer Assisted Qualitative Data Analysis Software CAQDAS will be used to manage this data. Both qualitative and quantitative data will be made available through publications and through metadata tagged on-line datasets where possible.	For Theatre in Home data analytics, the logging and monitoring infrastructure used the Elastic Stack instance provided within the Mantl platform. This infrastructure enabled logs to be generated by all 2-IMMERSE services, as well as each Client Application (running on a TV emulator or companion device), to be time-stamped and aggregated using a single consistent logging format. Logs were viewed, analysed and interpreted using the Kibana web application. For MotoGP Rancher (see earlier reference) was used. We also planned to use Google Analytics as a complementary solution for logging of user interactions with DMApp Components. Unfortunately it was not possible to achieve this for the Theatre at Home trial within the time available. We used survey monkey analytics –for Theatre at Home online surveys. MotoGP trials– Acumen are using a bespoke survey app which allows for data to be collected offline (in participants homes) and then uploaded later. The survey software they use to power this is Limesurvey. MotoGP analytics are via Elastic Stack & Google Analytics.
2.1	Code specifications will be released through project deliverables. The 2-IMMERSE system architecture is built from a number of defined services, each scoped with specific roles and responsibilities. These services will be designed to scale elastically, running the required	Code has already been securely stored and backed up via Github.

Section	Promise	Did we do it?
	number of instances to meet dynamic load requirements.	
2.1	As described in D2.1 the project will use Mantl, http://docs.mantl.io/en/latest/ a modern platform for rapidly deploying globally distributed services. Mantl provides an integrated set of industry-standard open-source components. It is cloud infrastructure provider agnostic, and can be deployed on AWS, OpenStack, Vagrant, Bare Metal etc. Mantl is licensed by 2-Immerse partner Cisco under the Apache Version 2 License.	Apart from Theatre at home operations, we have switched to Rancher (see above) for many operational reasons
2.2	Our default intention is to make the results of the project available through Open Source repositories. We are exploring the use of OpenAIRE https://www.openaire.eu/ repository https://zenodo.org/ Zenodo.	To be done
2.3	One of the goals of the project is to provide a functioning implement HbbTV 2.0 and inform the next generation of that standard. Data management of the code created will be the responsibility of the technical team under Cisco's leadership and the user data management will be the responsibility of the social science team lead by BBC.	Code Data management is already being carried out using Github
2.4	Data when generated will be kept in secure encrypted repositories within the partners' organisations. when it is released for wider availability it will be through secure stores such as OpenAIRE on the CERN sponsored Zenodo facility.	To be done