

Directorate General for Communications Networks, Content and Technology
Innovation Action

ICT-687655



D1.3 Data management Plan

Due date of deliverable: 31 May 2016

Actual submission date: 28 Sep 2016

Start date of project: 1 December 2015

Duration: 36 months

Lead contractor for this deliverable: **BBC**

Version: **25 Aug 2016**

Confidentiality status: **Public**

Abstract

This document is the first iteration of the project's Data Management Plan.

Target audience

This is a public deliverable and could be read by anyone with an interest in the data sets generated by 2-IMMERSE. This document will be read by the Project Consortium as it generates and collects data throughout the project.

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: WP1 Project Management

Document title: D1.3 Data Management Plan

Editor: Phil Stenton, BBC

Workpackage Leader: Phil Stenton, BBC

Project Co-ordinator: H el ene Waters, BBC

Project Leader: Phil Stenton, BBC

This project is co-funded by the European Union under the Horizon2020 research programme.

Copyright notice

  2016 Participants in project 2-IMMERSE

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 other 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.

List of Authors & Reviewers

Author: Phil Stenton - BBC

Reviewers: Ian Kegel – BT & H  l  ne Waters – BBC

Table of contents

Executive Summary	3
List of Authors & Reviewers	4
Table of contents.....	5
1. Data Summary.....	6
1.1. Purpose	6
1.2. Re-use.....	6
1.3. Origin of the Data.....	6
1.4. Expected Size and Utility	7
2. Fair data.....	7
2.1. Making data findable, including provisions for metadata	7
2.2. Making data openly accessible.....	8
2.3. Making data interoperable.....	8
2.4. Data security.....	8
2.5 Ethical aspects	8

1. Data Summary

Data will be collected during experiments and pilots (field trials) 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 address the general user questions that straddle the 4 pilots. 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.

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 composable micro-services called Distributed Media Applications, clusters of which may enable people to define their preferred media experience e.g. layout and synchronization of media across multiple and disparate devices.

1.2. Re-use

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, duvet 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 attending the 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 recordings of video chat and observation. This behavioural data may be useful to anyone wishing to understand the potential for multi-screen 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 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. An 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 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.

2.2. Making data openly accessible

The data produced and/or used in the project which 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](#) repository [Zenodo](#).

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 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 led 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 will be kept when generated 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

and 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.