Document Title: D1.2 ORDP: KPLEX - Open Research Data Pilot – 2018-01-31
Version: 0.2

Project Acronym:	KPLEX
Project Title:	Knowledge Complexity
Funding Scheme:	H2020-ICT-2016-1
Grant Agreement Number:	732340
Plan ID:	732340
Principal Investigator:	Jennifer Edmond
Principal Investigator ID:	0000-0001-9991-1637
Plan Data Contact:	Michelle Doran (doranm1@tcd.ie)
Project Start Date:	01 January 2017
Project End Date:	31 March 2018
Description:	Data Management Plan lists all the data that will be collected, processed and generated, within the project and details all rules, regulations and provisions connected to handling this data.

K•PLEX

KNOWLEDGE COMPLEXITY

	Dissemination Level	
P	Public	
C	Confidential, only for members of the consortium and the Commission Services	X

Revision History

Revision	Date	Author(s)	Description
0.1	21.08.2017	KPLEX Project Team	Complete Draft
0.2	31.01.2018	KPLEX Project Team	Final Version

Contents

Abbreviations

DANS	Data Archiving and Networked Services
DC	Dublin Core
DDI	Data Documentation Initiative: https://www.ddialliance.org/
DMP	Data Management Plan
FUB	Freie Universität Berlin
HAL	HAL open archive: https://hal.archives-ouvertes.fr/
OAIS	Open Archival Information System
ORDP	Open Research Data Pilot
PIL	Participant Information Leaflet
TCD	Trinity College Dublin
WP	Work Package

Executive Summary

This Data Management Plan (DMP) describes the elements of the data management life cycle of the Knowledge Complexity (KPLEX) project. The details included herein will continue to be developed in-line with the project's progress towards its objectives. The DMP consists of a living document created in the templates within the 'DMPonline' tool: part of the Open Research Data Pilot (ORDP) funded under Horizon 2020.

KPLEX is funded under the European Commission's Horizon 2020 research programme to undertake a 15-month investigation of the ways in which a focus on 'big data' in ICT research elides important issues about the information environment that we line in. KPLEX is part of the Open Research Data Pilot, a flexible pilot running under Horizon 2020 which aims to improve and maximise access to and re-use of research data generated by projects. Consequently, the DMP will provide information of how the research data of the project follows and adheres to the FAIR-principles of data management.

The Data Management Plan will provide information on the following elements in the data management lifecycle as specified in the 'DMPonline' tool:

- Type of data collected, processed and generated by the project
- The ways the project data complies with the FAIR-principles of data management
- Allocation of project resources
- Data security provisions
- Ethical rules of data management

The final dataset will be archived with Data Archiving and Networked Services (DANS) as a single collection using the online archive system EASY. EASY has been certified according to the guidelines of the Data Seal of Approval (DSA), the World Data System (WDS) and Nestor Seal for Trustworthy Digital Archives.

The Project Coordinator is Jennifer Edmond (Trinity College Dublin) edmondj@tcd.ie.

1. Data Summary

The purpose of the data acquisition is to query the attitudes and practices of current researchers in the areas of humanities and cultural sciences data and dealing with data complexities. Details of the digital outputs of this project are as follows:

- When speaking of documents generated within the KPLEX project, a distinction has been made between project documents that are deliverables as specified in the Grant Agreement and project documents that are peer-reviewed scientific publications.
  A. Project Documents (deliverables that are reports, reports, other documents)

  Every Work Package has a number of deliverables attached to it. Some of these deliverables are reports and vital to understanding the research project as a whole. Final deliverables as specified in the Grant Agreement will be deposited as part of the final dataset.

  B. Project Documents (peer-reviewed scientific publications)

  Final versions of peer-reviewed scientific publications will not be included in the archived dataset. However, the project team is committed to ensuring open access (free of charge, online access for any user) to all peer-reviewed scientific publications relating to its results within twelve months of publication.

- Audio files will be generated from face to face interviews with a sample of humanities and cultural science researchers, computer science researchers, emotion researchers in various disciplines and cultural heritage professionals.

- Qualitative and quantitative data will be generated from face-to-face interviews and online surveys and questionnaires with a sample of humanities and cultural science researchers, computer science researchers, emotion researchers in various disciplines, cultural heritage professionals and relevant policy makers as well as software developers and representatives of research funding bodies and private companies. According to the DANS deposition instructions for EASY, data files in the Social and Behavioural Sciences should contain clear and complete variable labels and value labels. Essential documentation files are:

  - Questionnaires or other research instruments

  - The fieldwork report (if available)
  - A code book, or a description of variables and information regarding (if applicable):
    - The population
    - Type of data (units of observation / analysis)
    - The sample and the sampling procedure
    - Response and non-response
    - The data collection method
    - Weighting variables
    - Constructed and/or derived variables
    - Information on anonymizing
  - Publications based on the data (if available) or a bibliographical description of such publications.

- Quantitative data will also be generated from a text mining exercise of publications retrieved from the ACM Digital Library and the Sage Journals Digital Library. For the text mining exercise, the dataset comprises of:

  - (1) text retrieved from full text of publications retrieved from the ACM Digital Library and the Sage Journals Digital Library (.pdf)
  - (2) converted .txt files
  - (3) the results of the text retrieval/ text mining exercise conducted on these .txt files.

- A project website built in Wordpress and located on Wordpress servers. Website content generated over the course of the project will be deposited as part of the KPLEX archived dataset.

- It is possible that the data may be used in future research of a similar nature by the research team and where possible, project data will be made available for reuse by scholars from outside of the project so as to support the project team's commitment to the promotion of Open Science and FAIR Data Principles.

Output #	WP#	Digital Output	Type	Format/Duration/ Size	Planned Access
1	1-6	Project Documents	Text	PDF/A (text, approx. 5MB)	Open Access via EASY at DANS as part of the KPLEX dataset & HAL open archive
2	2-4	38x interviews	Digital Audio	MP3, approx. 60 min each, 1.8GB	Project team
3	2-4	38x interview transcripts	Text	Anonymised Unicode text (.txt)	Open Access via EASY at DANS as part of the KPLEX dataset
4	2	Text retrieved using text mining	Text	Unicode text (.txt)	Project team
5	2-4	Questionnaires and interview questions	Text	PDF/A (text, approx. 5MB)	Open Access via EASY at DANS as part of the KPLEX dataset
6	3-4	Survey results	csv/sav	Anonymised SPSS (.sav) or data (.csv) + setup (.txt)	Open Access via EASY at DANS as part of the KPLEX dataset
7	2-5	Codebook	Text	Unicode text (.txt)	Open Access via EASY at DANS as part of the KPLEX dataset
8	2-4	Computer Assisted Qualitative Data Analysis (CAQDAS)	Text	ATLAS.TI copy bundle; NVIVO export project PDF/A (text, approx. 5MB)	Open Access via EASY at DANS as part of the KPLEX dataset
9	1-5	Website and blog posts	Text/ HTML	HTML (.html) plus related files: .css, .xslt, .js, .es as appropriate	Open Access via EASY at DANS as part of the KPLEX dataset

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- The final dataset will be archived with Data Archiving and Networked Services (DANS) as a single collection using the online archive system EASY. EASY has been certified according to the guidelines of the Data Seal of Approval (DSA), the World Data System (WDS) and Nestor Seal for Trustworthy Digital Archives. The metadata fields in EASY comply with the guidelines of the Dublin Core standards.
- Final versions of peer-reviewed scientific publications will not be included in the archived dataset. However, the project team is committed to ensuring open access (free of charge, online access for any user) to all peer-reviewed scientific publications relating to its results within twelve months of publication.
- DANS EASY automatically generates a *Persistent Identifier* in every new dataset submission: this is a unique identification with a permanent link that will always point to the dataset.
- Keywords will be added to the metadata of the deposited datasets, and all documents. The keywords will always include the following: KPLEX-Project, Horizon2020, H2020-ICT-2016-1.
- Over the course of data collection, a clear versioning system will be used consisting of file naming conventions (ProjectDescription-Date yyyy-mm-dd), together with standard headers listing creation dates and version numbers. Only final versions will be deposited as part of the archived datasets.
- Data Archiving and Networked Services (DANS) will provide facilities for long-term preservation of digital outputs created during the period of the KPLEX project in accordance with DANS archiving protocols and procedures for a minimum of 10 years after completion of the project.

2.2 Making data openly accessible:

- The digital output of the project will be archived with Data Archiving and Networked Services (DANS) as a single collection using the online archive system EASY. This dataset — with the exception of material containing sensitive variables — will be made available to download via the EASY interface. EASY archives are free to use for registered users.

- As it is technically difficult to anonymise interview audio files, these will not be openly available. Instead files will be archived securely with a restricted access policy applied.
- Redacted and anonymised transcripts of the audio files will be archived and made available in accordance to DANS protocols for anonymized data.
- Sensitive variables within the archived dataset will be stored at DANS in an encrypted spreadsheet which will be accessible by the project team for a minimum period of 10 years.
- Archiving of the spreadsheet, transcriptions and audio files will be undertaken in accordance with the DANS protocols for sensitive interview data.
- The results of the data mining investigation will not be available for sharing under the ORDP, as the original sources will have been accessed via the Trinity College Dublin subscriptions to the journals in question.

2.3 Making data interoperable:

- The final dataset will be archived with DANS using the online archive system EASY, with metadata compiled to their standards, based on DC terms.
- The team will investigate an appropriate standard vocabulary for all data types to allow for inter-disciplinary interoperability.
- The EASY interface will present the data in open formats enabling wider re-use.

2.4 Increase data re-use (through clarifying licenses):

- The project team proposes to apply the appropriate creative commons licences to the project data. This will reflect the different data types. A final decision will be made when the data is deposited.
- It is anticipated that the data will be made available in the third quarter of 2018.
- Data will be archived according to the guidelines of the international Data Seal of Approval, the ICSU-WDS, and the Nestor seal for Trustworthy Digital Archives. The Netherlands Code of Conduct for Scientific Practice (VSNU, 2014, in Dutch) prescribes a minimum retention period of ten years for raw data.

- DANS does not consider ten years to be “the long term”; its oldest available data date back as far as 1964. In other words, after the minimum retention period your data will remain accessible in the sustainable archives.

3. Allocation of resources

- Data management will be overseen by Trinity College Dublin together with the individual beneficiary team leaders during the data collection phase, and latterly by DANS in accordance with the international reference model for an Open Archival Information System (OAIS).
- Costs for acquisition, processing and analysis of the data, as well as those related to documentation of the data are covered by the grant (European Union’s Horizon2020 research and innovation programme).
- The financial costs for ensuring management and presentation of the project dataset by DANS have been included in the original project design.

4. Data security

4.1 Data recovery, secure storage and transfer of sensitive data

Data security will be addressed for the period of data collection and analysis by the project partners (1) and the deposition of the archive with DANS (2)

1) During the lifetime of the project, data collected from the face-to-face interviews, online surveys and questionnaires will be stored on the laptops of the researchers directly involved in the project. These laptops are password-protected as per local institutional requirements. In addition, data will be stored on password-protected hard drives of members of the project team. Upon completion of data collection, data will be transferred to an instance of DataverseNL (at DANS).

Data will be collected and stored in flexible and standard format (to avoid later loss of data through degradation of proprietary formats) and back up the data using a ‘321’ protocol (at least three copies of the data stored on at least two different media with at least one copy stored off site). As it is technically difficult to anonymise interview audio files, these will be stored securely with a restricted access policy applied. A unique code will be applied to anonymise the participant's details when used in any openly accessible outputs from this project including the interview transcripts. These data will be removed as soon as the data is successfully archived.

The approach is in full compliance with the policy of the coordinating institution for research of this nature.

2) At the end of the project the final dataset will be archived with Data Archiving and Networked Services (DANS) as a single collection using the online archive system EASY for a minimum retention period of 10 years after completion of the project. All data is stored on servers within the Netherlands.

Redacted and anonymised transcripts will be archived and made available in accordance to DANS protocols for anonymized data. Sensitive variables within the archived dataset will be stored at DANS in an encrypted spreadsheet which will be accessible by the project team for a minimum period of 10 years.

The plans and procedure regarding crises are laid down in the Business Continuity Plan –Crisis management plan DANS 2015 (available on request). The document (in Dutch) includes information on:

- The composition, roles and responsibilities of the crisis management team;
- The protocol when a crisis occurs;
- Information on the susceptibility of the relevant locations for the operation of the data archive.

4.2 What are the privacy issues that concern the acquisition of the data, if any?

In the case of fact-to-face interviews, a 'Participant Information Leaflet' (PIL) will be provided to potential participants in advance of any participation in the project. This leaflet will include information under the following headings: About; the KPLEX Project; What we want you to do; What happens if you can’t or don’t want to finish the interview; What if you change your mind after the interview; What happens to your data; Conflicts of Interest and Illicit Activities. After receiving this information and being given an opportunity to read it and ask any questions they might have, participants will then be asked to sign a consent form prior to their initial participation in this study. They will keep a copy of this signed form, and a copy will also be kept by the interviewer in a secured office / filing cabinet. Participants will be given a 7 day 'cooling-off' period, during which they can request that their data be withdrawn from the study (which includes any recordings) with no further consequences.

In the case of the online surveys, information of a similar nature to that in the full PIL will be provided to potential participants before they commence any participation, including the objectives of the project, the fact that their participation is voluntary, that data gathered from the survey will be securely stored for a period of time after the survey and the fact that the survey results will be fully anonymised and may be shared beyond the project.

5. Ethical aspects

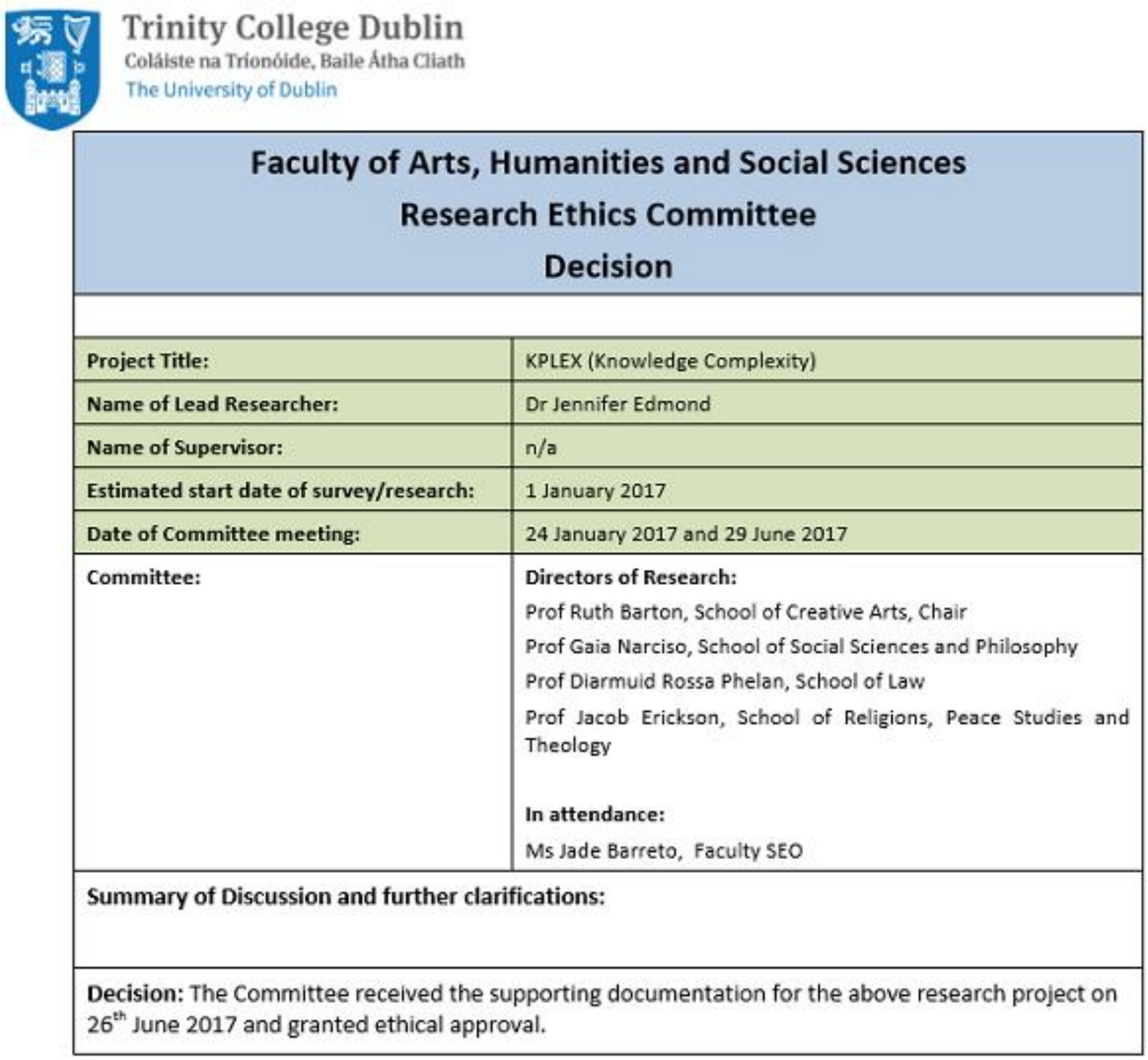
All ethical aspects connected to data collection and data sharing are overseen by an external ethics expert, namely the Chair of the Trinity College Dublin Faculty of Arts, Humanities and Social Sciences Ethics Committee (as nominated by Faculty Dean Darryl Jones).

The Ethics Committee of the Faculty of Arts Humanities and Social Science within Trinity College Dublin is comprised of a Chair (nominated annually by the Dean of Faculty of Arts, Humanities and Social Sciences) and 5 Faculty Representatives – normally Directors of Research – with one of the members always to be from the School of Law.

The members of this Committee are tasked: to read and review relevant applications, with a view to identifying potential ethical issues and thereby granting or refusing consent; to detail any concerns regarding the application from an ethical perspective; to review any notification of adverse effect; to review any monitoring reports submitted to the Committee; to review any end of project reports submitted to the Committee; and to keep up to date with legislative / best practice guidelines in the Ethics arena. More information about this Committee is available at:
https://ahss.tcd.ie/Faculty%20Ethics%20Committee/FacultyEthicsCommittee.php

A copy of the Trinity College Dublin, Faculty of Arts, Humanities and Social Science Research Ethics Committee decision can be found in Annex I.

Annex I: Faculty of Arts, Humanities and Social Science Research Ethics Committee Decision



Trinity College Dublin
Coláiste na Trionóide, Baile Átha Cliath
The University of Dublin

Faculty of Arts, Humanities and Social Sciences
Research Ethics Committee
Decision

Project Title:	KPLEX (Knowledge Complexity)
Name of Lead Researcher:	Dr Jennifer Edmond
Name of Supervisor:	n/a
Estimated start date of survey/research:	1 January 2017
Date of Committee meeting:	24 January 2017 and 29 June 2017
Committee:	**Directors of Research:** Prof Ruth Barton, School of Creative Arts, Chair Prof Gaia Narciso, School of Social Sciences and Philosophy Prof Diarmuid Rossa Phelan, School of Law Prof Jacob Erickson, School of Religions, Peace Studies and Theology **In attendance:** Ms Jade Barreto, Faculty SEO
Summary of Discussion and further clarifications:	
Decision: The Committee received the supporting documentation for the above research project on 26th June 2017 and granted ethical approval.	