D1.3 Data Management Plan

Ethan McCutchen ▸ Decko Commons ▸ 7/22/2016

Dissemination level	Public
Contractual date of delivery	Month 6 | June 2016
Actual date of delivery	Month 7 | July 2017
Work package	WP1 | Project Management
Deliverable number	D1.3 | Data Management Plan
Type	Report
Approval status	Approved
Version	3.2
Number of pages	17
File name	D1_3-20160722_V32_DC_Data_Management_Plan.docx

Abtract
To ensure reporting compliance with the Horizon 2020 Open Research Data pilot action, ChainReact datasets have been identified and herewith disclosed and detailed including criteria of selection, handing, methodology, lifecycle, and process.

Co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	2016-07-01	Outline	Ethan McCutchen
1.0	2016-07-11	Methodology, templates, and first datasets	Ethan McCutchen
1.1	2016-07-14	Whistle-related datasets	Richard Mills
1.2	2016-07-14	CERTH-managed datasets	Sotiris Diplaris
2.0	2016-07-15	General editing and addition of abstract and executive summary	Hala Khalaf , Ethan McCutchen
2.1	2016-07-18	Datasets managed by WikiRate eV	Vishal Kapadia
2.2	2016-07-20	Datasets managed by OpenCorporates	Chris Taggart
3.0	2016-07-20	Conclusion and formatting	Hala Khalaf , Ethan McCutchen
3.1	2016-07-21	Additional Whistle-related dataset	Richard Mills
3.2	2016-07-22	WP7 datasets	Łukasz Jonak

Author list

Organization	Name	Contact information
Decko Commons eV	Ethan McCutchen	ethan@decko.org
Cambridge University	Richard Mills	rm747@cam.ac.uk
WikiRate eV	Hala Khalaf	hala@wikirate.org
WikiRate eV	Vishal Kapadia	vishal@wikirate.org
CERTH	Sotiris Diplaris	diplaris@iti.gr
OpenCorporates	Chris Taggart	chris.taggart@opencorporates.com
DELAB	Łukasz Jonas	lukasz@jonak.info

Table of Contents

1 Executive Summary

In accordance with the European Commission Directorate-General for Research & Innovation "Guidelines on Data Management in Horizon 2020" v.2.1, the ChainReact Consortium partners collected, analysed and selected a series of datasets that corresponded to their progress regarding ChainReact's three main struts: The Whistle, OpenCorporates, and WikiRate. Each Consortium partner received a call-to-action to introduce the datasets most relevant to their respective deliverables.

This document reflects on the current state of Consortium agreements on the datasets that are produced and managed and outlines these sets of data in detail in terms of their description, selection methodology, use, owner, effect, data sharing principles and agreements, and lifecycle.

The data management plan will remain alive and evolving throughout the lifespan and the project. A second submission of the DMP will take effect at month 12. The datasets may also be altered due to converging factors such as project maturity, shifts in consumer usage, shifting to following working phase, etc.

2 Methodology

The methodology followed for drafting this initial DMP adheres to the European Commission's Guidelines[1] as interpreted in the online tool DMPonline[2]. DMPonline produced by the UK's Digital Curation Centre (DCC)[3] to help research teams address DMP requirements by addressing a series of questions for each dataset a project produces.

Accordingly, ChainReact's Initial DMP addresses the fields below for each dataset:

- Data set reference and name
- Data set description
- Standards and metadata
- Data sharing
- Archiving and preservation (including storage and backup).

2.1 Dataset reference and name

This field is the identifier for the dataset to be produced. The ChainReact dataset identification follows the naming: Data_<*WPno*>_<*serial number of dataset*>_<*dataset title*>. Example: **Data_WP2_1_Wikirate_Site**.

2.2 Dataset description

In this field the data that will be generated or collected is described, including references to their origin (in cases where data iare collected), nature, scale, to whom it could be useful, and whether it underpins a scientific publication. Where applicable, information on the existence (or non-existence) of similar data and the possibilities for their integration and reuse are mentioned.

2.3 Standards and metadata

This field examines existing suitable standards within relevant disciplines, as well as an outline on how and what metadata will be created. The available data standards (if any) accompany the description of the data that

[1] European Commission, (16 December 2013), *Guidelines on Data Management in Horizon 2020*, Version 1.0
[2] https://dmponline.dcc.ac.uk/
[3] http://www.dcc.ac.uk/

will collected and/or generated, including the description on how the data will be organised during the project, mentioning for example naming conventions, version control and folder structures.

The DCC provides the following questions to be considered as guidance on Data Capture Methods:

- *How will the data be created?*
- *What standards or methodologies will you use?*
- *How will you structure and name your folders and files?*
- *How will you ensure that different versions of a dataset are easily identifiable?*

2.4 Data sharing

In this field we describe how data will be shared, including access procedures, and embargo periods (if any). We also outline the technical mechanisms for dissemination, including necessary software and other tools for enabling re-use; define the breadth of access.

In case the dataset cannot be shared, the reasons for this will be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

2.5 Archiving and preservation

Here the procedures that will be put in place for long-term preservation of the data will be described, along with the indication of how long the data should be preserved, what is its approximated end volume, including a reference to the associated costs (if any) and how these are planned to be covered. This point emphasizes in the long-term preservation and curation of data, beyond the lifetime of the project. Where dedicated resources are needed, these should be outlined and justified, including any relevant technical expertise, support and training that is likely to be required and how it will be acquired.

3 ChainReact Datasets

3.1 WP1 Datasets

Data set reference and name	**Data_WP1_1_ChainReact_Docs_Site**
Data set description	A restricted Wagn-based website at docs.chainreact.org used for internal collaboration of all ChainReact partners. Will include the canonical versions of reports, deliverables, proposals, and core results of huddles and other meetings. Because of the flexibility of this platform, it will often be used for creating structures for organizing other data collaborated on by many partners.
Standards and metadata	Like all Wagn sites, the docs site is organized into "cards". For every edit of every card (including name, type, and content changes), wagn stores: • a userstamp • a timestamp, and • an IP address. When multiple cards are edited simultaneously, these independently tracked "actions" are grouped into single "acts". It is also possible to collect additional metadata and standards-conforming data within cards.
Data sharing	By default, cards on the docs site are restricted to viewing by partners, though any individual card may be independently made publicly viewable if deemed appropriate by its editors. Much of the site's content is material being prepared for publication but not appropriate for publication in raw states. Other cards contain conversations, personal data, and proposals that have been rejected or not yet agreed upon. It is, by and large, a site for process rather than final products.
Archiving and preservation	The docs site is currently stored on the WikiRate production server and will likely be moved to a smaller server when WikiRate.org moves to a multi-server architecture. Full site backups are automatically generated daily, with one copy stored locally and another transferred to our development server. Wagn automatically handles card revisions, and the complete history of every card is visible via the interface. Decko Commons eV has accepted responsibility for continued hosting of and updated to the website after the project's completion. Should it be unable to continue hosting at some point in the future, it will provide all partners with an archive, which will be made conveniently usable with the installation of the open-source Wagn/Decko platform.

Data set reference and name	**Data_WP1_2_Contacts_Database**
Data set description	Lists of key Contacts at partners, hosted in the form of mailing lists
Standards and metadata	Standard form of name, and email address, organised into general and project specific mailing lists (e.g. financial contacts, WP coordination)
Data sharing	These contact lists are viewable by the ChainReact project team and editable by the administrators, at WikiRate e.V.
Archiving and preservation	The data is stored and maintained in ChainReact's Google apps account

3.2 WP2 Datasets

Data set reference and name	**Data_WP2_1_Whistle_Research_Informing_Design_Data**
Data set description	This data-set includes all data collected in relation to research that informs the design of The Whistle. The nature of this data will include audio-visual recordings of interviews and user testing sessions - along with the associated consent forms, transcriptions, interview/test plans and participant recruitment lists/documents. This data-set will be stored in a google drive folder, and relevant people from the project team will be granted access. This data-set is likely to support scientific publications, in which case transcripts or excerpts may be shared alongside these publications. This data-set will not be particularly large, and should not exceed 1 gigabyte in size.
Standards and metadata	The top-level google drive folder will contain sub-folders for the following: • Documents containing interview questions and related materials • Interview recordings • Interview transcripts • Interview recruitment tracking Files relating to interviews will be stored within sub-folders named for the organisation they relate to with titles denoting the person who was interviewed.
Data sharing	This data-set will be shared with all relevant project team members through their google accounts.
Archiving and preservation	As this data-set will be stored in a google drive folder, it will benefit from a version history and there should be no issue with its preservation.

3.3 WP3 Datasets

Data set reference and name	**Data_WP3_1_Whistle_Reports**
Data set description	A restricted data-set encompassing the full detail of all incoming civilian witness reports and attachments for The Whistle. The Whistle will run reporting campaigns, in collaboration with NGOs, to collect reports from civilian witnesses. When a civilian witness submits a report this will create a record on The Whistle's secure server, for the purpose of the data management plan all such reports are being treated as a single data-set. In practice, only nominated representatives of the partner NGO for each campaign will be allowed to access reports related to that campaign. The precise nature and scale of this data-set will depend on the choice of reporting campaigns. Ethics deliverable 9.2 contains further detail on how this data will be stored and transmitted, and deliverable 2.1 contains detail on the ethical review of prospective campaigns (which includes review of which data will be stored and procedures for data collection). This data-set will contain sensitive information, and therefore storing and transmitting it securely is a central concern for the project. This data-set may be used in academic research, and therefore may underpin a scientific publication.
Standards and metadata	As The Whistle is in the early stages of development the choice of a specific standard for storage of this data is yet to be made. The data for incoming reports will be similar to that produced by standard web forms that allow attachments. The choice of a specific standard will be determined by security considerations. When a report is submitted, it will be stored along with meta-data such as the time of creation and IP address of submitter. The Whistle will also allow aspects of a report to be passed through relevant external APIs that could facilitate work on verification of its authenticity. Results of these API calls will also be stored as additional meta-data for a report.
Data sharing	Due to the sensitive nature of this data-set, access will be tightly restricted. Only nominated representatives of the partner NGO for a campaign, and relevant people within the project team, will have access to this data. A reporting campaign may also produce aggregated or de-personalised data that can be published on sites like wikirate.org (thus forming part of the Data_WP5_1_WikiRate_Site_Cards data-set). The manner in which public-facing data is produced for a campaign will be considered as part of the ethical review for a prospective reporting campaign.
Archiving and preservation	The full data for reports will be retained on the secure server until 3 months after the reporting campaign ends - at which point it will be transferred to a secure archive housed separately to the data for live campaigns. Data held in this archive will only be used for research purposes. Preservation of this archived data-set will be the responsibility of the research team at Cambridge. At the point when this data-set serves no further research purpose, or cannot be maintained securely, it will be destroyed.

3.4 WP4 Datasets

Data set reference and name	**Data_WP4_1_Possible NGO Partners**
Data set description	Contacts and engagement data-set to track charities that could be partner with The Whistle to run test reporting campaigns
Standards and metadata	Data is stored in a Google Sheet with columns representing: • Charity name • Location • Website • Contact Email • Funding Band • Purpose • Digital Literacy • Country Focus • Population Focus • Notes • Interview Status
Data sharing	This data set will be shared with all relevant team members working on the interview study and outreach with possible partners for The Whistle.
Archiving and preservation	As this data-set is stored in a google sheet it will benefit from a version history and there should be no issue with its preservation.

3.5 WP5 Datasets

Data set reference and name	**Data_WP5_1_WikiRate_Site_Cards**
Data set description	The primary Wagn database for the WikiRate.org website. (Note that the assets for this website are treated as a separate dataset, because they will involve separate archiving and preservation.) All of WikiRate's core concepts – Companies, Metrics, Topics, Claims, Reviews, Sources, and Projects – as well as more standard content like Users and simple webpages, are organized as cards within a wagn website.
Standards and metadata	Like all Wagn sites, WikiRate.org is organized into "cards", and all data are stored in the same five tables (cards, card_acts, card_actions, card_changes, and card_references.) As noted in *Data_WP1_1_ChainReact_Docs_Site* above, for every edit of every card (including name, type, and content changes), Wagn stores: • a userstamp • a timestamp, and

Data set reference and name	**Data_WP5_1_WikiRate_Site_Cards**
	• an IP address. Wagn also supports a REST API that allows this data to be made available in many formats. Company data will be made available in many standard formats, including JSON, XBRL, and simpler formats like CSV. Many metrics themselves contain standardized data. Initially, standards conformity will be enforced via community feedback and editing, though some automation will likely be added in later stages.
Data sharing	Account login information, including encrypted passwords, are protected and made invisible to web users. All other information on WikiRate.org is available for reading and download by the general public. Some metric data providers have requested download limitations so that their original datasets could not be reconstructed from WikiRate.org. We are currently weighing the benefits of supporting such limitations (and thus receiving permission to put more data on WikiRate.org) vs. the costs of having to support more restrictions and communicate the nature of and rationale for these restrictions to users.
Archiving and preservation	Development and promotion of this dataset is the core focus of WikiRate eV, who intend to see it thrive and grow long after the end of the current project, supported by broad fundraising and community-building strategies. The entire database is archived nightly, with a full version tarred and copied to a remote server. We also frequently make full and partial copies to various servers for use in development and testing. Some site copies are used for experimenting with data that we are not yet ready to publish for technical or social reasons, most commonly permission not yet granted. Wagn automatically handles card revisions, and the complete history of every card is visible via the interface.

Data set reference and name	**Data_WP5_2_WikiRate_Site_Assets**
Data set description	Files uploaded to WikiRate.org, including images, structured and unstructured source files, and optimized CSS and JavaScript.
Standards and metadata	Metadata for these files are stored as cards in the previous dataset, *Data_WP5_1_WikiRate_Site_Cards*. Each asset is stored with a card_id and action_id that allows it to be mapped to that dataset. However, because our multi-server architecture calls for a canonical database engine on one server and canonical file service elsewhere, these two datasets will be tracked separately.
Data sharing	All files are publicly available. Direct links to the data are provided on WikiRate.org
Archiving and preservation	At present, the data remain on our production server and, like the production

Data set reference and name	**Data_WP5_2_WikiRate_Site_Assets**
	database, are archived and backed up nightly. Soon they will be moved to an independent server or cloud service in support of WikiRate.org's designed multi-server architecture. As with *Data_WP5_1_WikiRate_Site_Cards,* maintenance and development of this dataset is connected to the primary focus of WikiRate e.V. and will be central to ongoing planning, fundraising, and promotion.

Data set reference and name	**Data_WP5_3_CERTH_ Companies**
Data set description	CERTH's company entities collection that have been and are going to be obtained by Web data extraction using easIE (an easy-to-use information extraction framework).
Standards and metadata	A schema-free document-oriented database is used which allows us to add or remove fields from the collection without impact to the database soundness. Each company is described by the following: • id • company_name • aliases • website • address • country • wikirate_id: this field is present only in companies that have been integrated to WikiRate platform. • opencorporates_id: this field is present only if there is a matching entity in OpenCorporates database. Company mapping task will result to the integration of companies between OpenCorporates and WikiRate. Additional fields might be considered in order to represent the relationships between companies in our dataset derived from OpenCorporates corporate networks.
Data sharing	A RESTful API will be available for anyone who wishes to have access to the dataset. The data will be available in JSON format.
Archiving and preservation	Preservation will be ensured by backup of the original database.

Data set reference and name	**Data_WP5_4_Metrics**
Data set description	CERTH's metrics collections that have been and are going to be extracted from external Web sources by using easIE (an easy-to-use information extraction framework).
Standards and metadata	A schema-free document-oriented database is used which allows us to add or remove fields from the collection without impact to the database soundness. Each metric is described by the following:

Data set reference and name	**Data_WP5_4_Metrics**
	• name • value • referred_company • citeyear • source • source_name • type • currency
Data sharing	The collected metrics will be available through a RESTful API for anyone who wishes to have access to the dataset. The data will be available in JSON format. We encourage people and companies to reuse our data and contribute to data collection task regarding companies' CSR performance.
Archiving and preservation	Preservation will be ensured by backup of the original database.

Data set reference and name	**Data_WP5_5_WikiRate_Usability**
Data set description	Results of user testing and design, including think aloud tests, analytics, reading material, etc.
Standards and metadata	The top-level google drive folder contains sub-folders for the following: • Lean UX Activities • UX Design • UX Research Files will be named with descriptive titles coupled with date and version information. Interview recordings and transcript file names will contain the name of the organisation represented by the interviewee, a number denoting the interview's order and date information.
Data sharing	This data-set will be shared with all relevant project team members through their google accounts.
Archiving and preservation	As this data-set will be stored in a google drive folder, it will benefit from a version history and there should be no issue with its preservation.

Data set reference and name	**Data_WP5_6_OpenCorporates_Corporate_Relationship_Sources**
Data set description	This is the list of potential sources for relationship data, compiled for the report as part of WP5.1. This dataset is not kept in a database, but in the Google Doc, which is the master document for the report (rather than the derived Word Document supplied as a deliverable).
Standards and metadata	As this is kept in a Google Document, all changes to it are automatically tracked.
Data sharing	This is a list of "not yet published" data and is therefore private.
Archiving and preservation	As the dataset is in the cloud, there is automatic archiving. We also periodically export the report into different forms (e.g. Word Docs).

Data set reference and name	**Data_WP5_7_OpenCorporates_Companies**
Data set description	This dataset is the core dataset of over 100 million companies in OpenCorporates, all obtained from primary public sources by OpenCorporates
Standards and metadata	The OpenCorporates company data has multiple fields and attributes, often deeply nested and rich. The conceptual schema is described (using JSON-schema) at **https://github.com/openc/openc-schema/blob/master/build/company-schema.json** (this schema is open-source). All data is fully provenance, describing both the source and retrieval timestamp
Data sharing	The data is available through the OpenCorporates enterprise-level API (Application Programming Interface), which provides rich querying and retrieval.
Archiving and preservation	The data lives on our production MySQL database, which lives on our multiserver architecture (master + slave + backup slave), which is backed up daily, with historical backups.

Data set reference and name	**Data_WP5_8_OpenCorporates_Corporate_Structures**
Data set description	This dataset is the corporate structure information OpenCorporates has extracted from official public sources (includes shareholding, subsidiary, control relationships from company registers, SEC, other regulators)
Standards and metadata	The OpenCorporates corporate structure data is modelled using our own model, which is open source (see **https://github.com/openc/openc-schema/blob/master/build/** for schemas), and described in **a series of blog posts**. As the data comes from multiple sources, with varying levels of details and subtle differences in meaning (for example the way shareholding is represented), the models need to be able to cope with this, in particular both high and low granularity, significant ambiguities, and different natures of the relationship (e.g. shareholding, subsidiaries, other control relationships). All data is fully provenanced, describing both the source and retrieval timestamp
Data sharing	The data is available through the OpenCorporates enterprise-level API (Application Programming Interface), which provides rich querying and retrieval. As part of this project we will be working with the partners to enhance retrieval of corporate structure information via the API
Archiving and preservation	The data lives on our production MySQL database, which lives on our replicated multiserver cluster (master + slave + backup slave), which is backed up daily. In addition, we use a replicated Neo4J cluster for storing the relationships in a graph database. This is also backed up daily

3.6 WP6 Datasets

Data set reference and name	**Data_WP6_1_Corporate Engagement**
Data set description	Contacts and engagement database to help us identify targets and progression towards corporate engagement
Standards and metadata	Data will be collected through Google tracking sheets and where possible tracked in Salesforce software.
Data sharing	This data set will be shared with all relevant team members working on outreach, partnerships and engagement. Additionally analysis of this data set may be used at periodic project meetings to indicate progress and consider direction
Archiving and preservation	Salesforce is a dynamic database which will exist in perpetuity whilst WikiRate e.V. benefits from the non-profit license. If we ever need to migrate to another software the entire database can be exported. Google sheets will also exist in perpetuity, and offer a layer of tracking and analysis which Salesforce cannot capture alone.

3.7 WP7 Datasets

Data set reference and name	**Data_WP7_1_Collective_Awarness_Platforms_Research**
Data set description	This data set will be used to analyse the functioning of ChainReact as a prime example of Collective Awareness Platform. It will be the result of extract/transform process that will retrieve the data from various ChainReact databases (especially the repositories of The Whistle and Wikirate), combine them and transform into the form suitable for research. The dataset will describe in detail the actions of ChainReact users – their interactions with the platform, their uploads, their site navigation paths, etc. It will be used to calculate various indicators describing the overall functioning of ChainReact.
Standards and metadata	The specific technology and data model of research dataset is contingent on the final structure of source databases and the design of research the dataset will be used for. Both these aspects being under development, there is a wide range of storage choices being considered at the moment, from standard SQL schemas, to XML/JSON containers, to graph databases.
Data sharing	The data set will be initially shared among ChainReact members. It will be made available publicly as the background for research activities.
Archiving and preservation	The specific location of the dataset (and consequently archiving and preservation policies) is yet to be decided. Existing ChainReact infrastructure (servers) could be used or specific cloud or local solution be chosen, depending on the dataset and research requirements that will be decided in the course of the project.

Data set reference and name	**Data_WP7_2_Title_ChainReact_Evaluation**
Data set description	This data set comprises of various forms of data needed to evaluate ChainReact in terms of progress towards the realisation of its goals and the quality of its inner functioning. The data set includes the progress reports and other communication with consortium partners, the audio recordings and transcripts of interviews with ChainReact team members, participatory observation notes and results of desk research activities.
Standards and metadata	The data set will be stored as a google drive folder with subfolder structure reflecting the nature and structure of research material.
Data sharing	The data-set will be shared mainly among the researchers performing evaluation. The less sensitive elements of the data-set (eg. progress reports, desk research notes) will be made available for general reuse by Consortium, while one-to-one communication recordings will be treated as confidential and shared only among researchers directly involved in evaluation tasks. The access control to the data will be realised by google drive sharing mechanism with possibility of encrypting particular file containers as an extra security layer.
Archiving and preservation	The archiving of the data set will be realised by google drive persistence and versioning mechanism. The data-set will be stored for two years after final evaluation report, in accordance with evaluation standards, for the purpose of verification and auditing. After two years the data set will be discarded.

3.8 WP8 Datasets

Data set reference and name	**Data_WP8_1_NGO Engagement**
Data set description	Contacts and engagement database to help us identify targets and progression towards NGO engagement
Standards and metadata	Data will be collected through Google tracking sheets and where possible tracked in Salesforce software.
Data sharing	This data set will be shared with all relevant team members working on outreach, partnerships and engagement. Additionally analysis of this data set may be used at periodic project meetings to indicate progress and consider direction
Archiving and preservation	Salesforce is a dynamic database which will exist in perpetuity whilst WikiRate e.V. benefits from the non-profit license. If we ever need to migrate to another software the entire database can be exported. Google sheets will also exist in perpetuity, and offer a layer of tracking and analysis which Salesforce cannot capture alone.

4 Conclusion

This Data Management Plan identifies the datasets managed by the ChainReact consortium organized by work packages. As detailed under section 3 of this report "ChainReact Datasets", the nature of these datasets vary according to each components' roles and responsibilities. For example, CERTH's company metadata are collected and maintained through the easIE extraction framework and preserved through regular backup of the database, whereas the outreach plan set by WikiRate manages a database of contacts and leads that are categorised according to their outreach status (connection established/not, connection success/pending, etc.).

The ChainReact datasets are evolving. Therefore, DMP is a living document that will keep being updated through the lifetime of the project.