Grant Agreement No.: 761488

D6.2: Data Management Plan

Work package	WP 6
Task	T6.1
Due date	28/02/2018
Submission date	28/02/2018
Deliverable lead	VRT
Version	
Authors	Ferdinando Bosco (Engineering), Vincenzo Croce (Engineering), Christos Danezis (Dias), Zoë De Ruyck (imec), Daan Gerits (VRT), Iris Jennes (VUB), Olga Kisselmann (DW), Joris Mattheijssens (VRT), Natasja Van Buggenhout (VUB)
Reviewers	Chris Develder (UGent), Matthias Strobbe (UGent)
Keywords	CPN, Personalisation, personal data, data management, GDPR

Document Revision History

Version	Date	Description of change	List of contributor(s)
0.1 Initial version	14/02/2018	Questions answered for VRT, indications for other partners	Daan Gerits (VRT), Joris Mattheijssens (VRT)
0.2 Draft version	21/02/2018	Partners answered questions	Ferdinando Bosco (Engineering), Vincenzo Croce (Engineering), Christos Danezis (Dias), Olga Kisselmann (DW), Joris Mattheijssens (VRT)
1.0 Final version	28/02/2018	Adaptations required by reviewers made + extra information added.	Zoë De Ruyck (imec), Iris Jennes (VUB), Natasja Van Buggenhout (VUB)

Co-funded by the Horizon 2020
Framework Programme of the European Union

DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 761488.

Project co-funded by the European Commission in the H2020 Programme	
Nature of the deliverable:	**R**
Dissemination Level	
PU	Public, fully open, e.g. web
~~**CL**~~	~~Classified, information as referred to in Commission Decision 2001/844/EC~~
~~**CO**~~	~~Confidential to CPN project and Commission Services~~

Co-funded by the Horizon 2020 Framework Programme of the European Union

EXECUTIVE SUMMARY

The Data Management Plan explains how the different datasets of the CPN project are stored and shared with third parties.

The datasets are assembled and constructed in order to fulfill the project goals. The main datasets are the Article dataset, the User dataset, the User Behaviour dataset and the User Profile dataset. Possibly, datasets for NLP and social media will be added.

After pseudonymisation and aggregation of localisation data, the datasets will be made publicly available through an open repository. The thus shared datasets will comply to the GDPR and to the FAIR principles. Complying to these, makes the datasets both safe and ethically correct.

Efforts are underway to implement data management principles throughout all partners.

The DMP is a living document, which means that updated versions will be created as part of the progress being made in the project. This version reflects the status based on the information currently available in the project.

Co-funded by the Horizon 2020
Framework Programme of the European Union

TABLE OF CONTENTS

Co-funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

CSV	Comma Separated Value
DMP	Data Management Plan
DPO	Data Protection Officer
DW	Deutsche Welle
FAIR	Findable, Accessible, Interoperable, Reuseable
GDPR	General Data Protection Regulation
JSON	JavaScript Object Notation
NLP	Natural Language Processing
ORC	Optimized Row Columnar
VRT	Vlaamse Radio- en Televisieomroep
VUB	Vrije Universiteit Brussel

Co-funded by the Horizon 2020
Framework Programme of the European Union

1 DATA SUMMARY

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation.**

The purpose of the project is to create a better information distribution, while at the same time respecting user privacy at all cost, through personalisation of the media content. In order to achieve these goals, we will collect data on the reads of news articles, along with data on the content of the articles, and the structure of the front page. These datasets are essential for the identification of typical user interests and behaviour patterns. Possibly, secondary datasets from third parties will be linked to the identified patterns.

- **Explain the relation to the objectives of the project.**

Objectives:

1. to reach the next level of personalised and context aware recommendation services for news content, by significantly increasing the congruence between user, context, content and business/societal goals as a result of more advanced modelling and social engineering;
2. to provide media stakeholders with media-centered cross-platform distribution channels, including powerful services to gain actionable insights about their audience, consumption patterns and business goals allowing them to define and push actions such as adjusting their content strategy; and
3. to improve the user experience providing the media consumer with personalised news offering contextualised to their particular situation (preferences, time of the day, type of platforms, location, etc.). CPN will achieve this while especially taking care not to push the media consumer in a filter bubble.

The primary data contain information on the articles (title, publication date and time, text, …), on the users (platform, links, browser, localisation, …) and on both (which user opened which article when). These data are essential for the analysis of user behaviour. We want to be able to answer questions like:

- Which subgroups exist in the population of users?
- Are there typical interests and behaviours inside subgroups?
- Are there other factors that correlate to user interests and behaviour? Are there identifiable subgroups of articles?
- Is there a matching between subgroups of users and subgroups of articles? How can we use this knowledge to improve the content for an individual or for a subgroup of users?

- **Specify the types and formats of data generated/collected.**

Within the CPN project there will be different datasets, of different format, selected to offer a comprehensive analysis of the concept of media content consumption and to offer innovative services for content personalisation.
In particular:

- Article dataset: contains data and metadata of the articles published on the platform. Initially the source providers will be only the partners of the CPN consortium (VRT, DW, Dias) but in the future these could be more numerous (at least 8 external media organisations during the pilots)
- User dataset: contains user data, pseudonymised and compliant with the GDPR.

Co-funded by the Horizon 2020
Framework Programme of the European Union

- User Behaviour dataset : are the data collected based on user behaviour when using the CPN platform.
- User Profile Dataset: is an aggregation of different types of data related to the user, which are used to create customised services related to the contents offered on the platform

- **Specify if existing data is being re-used (if any).**

Possibly, we will re-use datasets for sentiment analysis and natural language (NLP) processing and datasets from social media scraping. For sentiment analysis and NLP, the University of Antwerp CLiPS Center offers an excellent database of annotated sentences and texts (https://www.clips.uantwerpen.be/pages/pattern-nl). The Python Natural Language Toolkit offers similar databases, which are already integrated in the Python programming language (https://www.nltk.org/). If social media scraping is found necessary for the project,, we will probably build our own databases.

- **Specify the origin of the data.**

The initial datasets will be provided by CPN media partners (VRT, DW, Dias), so as to have a sufficient database to initialise the platform. Then the data source will be the CPN platform itself.

- **State the expected size of the data (if known).**

At this point in the project it is not easy to quantify the size of the datasets . The platform could be initialised with a series of data (for example the last 60 days) extrapolated from those of the partners:

- Article Dataset: VRT + DW + Dias
- User Dataset : VRT + Dias +DW (to the extent stated in D4.1)
- User Behaviour Dataset: None (collected later)
- User Profile Dataset: None (collect later)

It will be possible to estimate the amount of data once the platform has been initialised.

- **Outline the data utility: to whom will it be useful**

The CPN platform will be open to third parties integration, so other researchers and developers in the field of media personalisation can use these data to develop models of user interests and behaviour, and to develop applications with personalised content.

Co-funded by the Horizon 2020
Framework Programme of the European Union

2 FAIR DATA

2.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

- **Outline the discoverability of data (metadata provision).**

The metadata will be published on an open archive portal, like Zenodo or git. Furthermore, all the datasets will be available and searchable within the CPN platform itself.

- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**

Using an open archive to publish dataset, CPN will guarantee:
- that a unique DOI is assigned to each data version
- that metadata will be discoverable through the portal search functionality and programmatically using a standard protocol (e.g. OAI-PMH).

- **Outline naming conventions used.**

CPN will use a naming conventions with the following information:
- Project name/acronym (CPN)
- Name of dataset
- Source provider
- Date file created/generated (in YYYY-MM-DD format)
- Version number

For example: CPN_articles_VRT_2018-02-20_v1.0.

The data will adopt the Semantic Versioning 2.0 scheme to assign a unique version to each release of the data.

- **Outline the approach towards search keyword.**

The goal of keywords is the facilitation of searches for datasets by third parties. The repository will contain several metadata files containing lists of keywords. We will provide at least one list of search keywords for each type of dataset. For example, for the article time series, the keywords “news”, “article”, “time series” and “chartbeat” are adequate search keywords. Also, the file format (CSV, JSON or ORC) is a useful keyword.

- **Outline the approach for clear versioning**

We will enable versioning to avoid unintended overwrites and deletions and to archive objects in order to retrieve previous versions of them. For example, when a file is deleted, it remains present, though hidden, with a special delete marker added to it.

One of the criteria for selecting an open repository provider, is that it has to offer an efficient and safe versioning system for these purposes.

Co-funded by the Horizon 2020
Framework Programme of the European Union

- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**.

The most important metadata files are those for the articles. We store information like the date and hour of publication, the title, the number of photos, the visibility on the website, and of course the content.

For example, for every new article that is published on the VRTnws website, a JSON string is added to a list in an endpoint. We poll a rolling window of the 500 most frequent articles at this endpoint every few minutes, creating and storing a new JSON file every time. Subsequent JSON files have an overlap, as there are never 500 new articles published between two polls.

For technical reasons related to the design of each news website, it is highly uncertain that all partners will poll their article metadata in the same way. Standardisation of the article metadata will happen one or two steps later, i.e. when the polled data have been preprocessed and prepared for storing in the open repository.

2.2 MAKING DATA OPENLY ACCESSIBLE

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so.**

The following datasets will be published in an open and completely accessible way:
- Article dataset
- User Behaviour dataset (anonymous)

Deutsche Welle will provide only a limited user behaviour dataset. DW user behaviour datasets must remain confidential at all times and only accessible for project purpose by CPN project partners. DW article-related datasets can be made publicly accessible to the extent, to which it was done in previous research projects.

The User Profile dataset contains information that could make users identifiable, like IP-addresses, genders, locations, etc. Therefore, it will be kept confidential.

- **Specify how the data will be made available.**

Once collected, the data will be aggregated daily and subsequently published on a public access repository.

- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**

All data generated in CPN are intended to be used by software and machines. Although it is possible to access and manually read raw data, it is generally not useful. Instead, it is recommended to make use of software for that purpose. While the CPN software can be used to read and evaluate the data, we will also ensure that the data sets and their formats are sufficiently well documented for others to use their own software for accessing them.

Co-funded by the Horizon 2020
Framework Programme of the European Union

- **Specify where the data and associate metadata, documentation and code are deposited.**

The datasets will be available on a publicly accessible repository, like Zenodo or git.

- **Specify how access will be provided in case there are any restrictions.**

Whenever possible, the datasets will be made available under a open license. Whenever the release of data conflicts with a partner's exploitation strategy (e.g. User profile dataset), the partner may pick a more restrictive license.

2.3 MAKING DATA INTEROPERABLE

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**

It is currently unclear to what extent data and metadata vocabularies, standards, and methodologies are applicable to data sets collected by CPN. One of the tools imec will work on during the project is the semantic annotation of the data. Therefore, it should become clear which vocabularies, standards, and methodologies will be used.

- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

It is currently unclear to what extent data and metadata vocabularies, standards, and methodologies are applicable to data sets collected by CPN. One of the tools imec will work on during the project is the semantic annotation of the data. Therefore, it will become clear in a later phase of the project which vocabularies, standards, and methodologies will be used.

2.4 INCREASE DATA RE-USE (THROUGH CLARIFYING LICENSES)

- **Specify how the data will be licensed to permit the widest reuse possible.**

We will strive towards a strict application of the GDPR when constructing and transforming the open datasets. Once the GDPR is fulfilled, we will allow access to the data to the widest audience possible.

- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed.**

Data will be made available for re-use as soon as possible. This is at the very earliest the case at the time of the submission of the respective deliverables. We expect that usually the release of the research data aligned with the deliverables will take around two months' time. In particular, an initial dataset will be release at the start of platform initialisation and successive datasets later after each pilot.

Co-funded by the Horizon 2020
Framework Programme of the European Union

- **Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.**

The data used in the project are usable by several third parties. For example, media researchers and innovators can use it for developing new models and innovative media services, and data scientists can train machine learning models with it. User data will however be restricted when sharing is not GDPR-compliant.

- **Describe data quality assurance processes.**

We will publish policies as to the level and type of their efforts concerning protection from data decay, alteration or loss, including what will happen to data the event of the repository going out of business. Users of the repository can judge the repository's methods and commitment to data integrity preservation. We will provide evidence that at least best efforts are made to protect the integrity of the data.

For file preservation, we rely on the services provided by the repository provider to preserve our data. All data files are stored along with a MD5 checksum of the file content. Files are regularly checked against their checksums to assure that file content remains constant.

- **Specify the length of time for which the data will remain re-usable.**

The data will remain reuseable for at least the duration of the project. We keep a 5 year retention policy on the data.

3 ALLOCATION OF RESOURCES

- **Explain the allocation of resources, addressing the following issues:**
  - **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs.**

CPN can not estimates at this time the cost for the transformation of the existing datasets into FAIR datasets.

- **Clearly identify responsibilities for data management in your project.**

The management of the data set is shared amongst the project partners who take part in contributing details about the datasets in the DMP deliverables, those that take care of the deposition of research data in open research data repositories, and overall through the Quality Manager.

- **Describe costs and potential value of long term preservation.**

CPN, through the use of an open archive, estimates no costs to maintain a FAIR dataset (e.g. costs for providing open access, related research data management costs, data curation and data storage costs).

Co-funded by the Horizon 2020
Framework Programme of the European Union

4 DATA SECURITY

- **Address data recovery as well as secure storage and transfer of sensitive data.**

Data recovery in case of loss will be possible through redundant storage of unprocessed data on an S3 server and processed data on an open repository. In order to be be GDPR-compliant, no sensitive data will be made available on the open repository.

5 ETHICAL ASPECTS

- **To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former.**

The GDPR becomes law on 25 May 2018. Since the pilots of CPN will be finalized after this date, and since all consortium partners are based in member states, except for Digital Catapult, the project must follow the GDPR. Digital Catapult is based in the UK, but non-EU countries dealing with the data of EU data subjects must also obey the rules of the GDPR. Furthermore, the UK will still be classed as a Member State on 25 May 2018.

CPN will be collecting and processing potentially sensitive personal data, but exclusively when it is necessary for the project. We will replace all personal identifiers, such as a name or IP-address, with a pseudonym as early as possible in the data chain. After pseudonymisation, no identifiable data will be stored.

Locations of users are a specific problem. We want to use the location of the users for the personalisation of (regional) news content. However, precise localisation of a user can sometimes, in combination with other data, be used to identify a user. We will protect localisation data in two ways:

- For internal use, we will make too precise localisations less precise, e.g. convert addresses (street and number) to large areas like provinces.
- For sharing with third parties, only aggregated data will be made available.

All consortium partners, that process personal data, have appointed or are planning to appoint a DPO (Data Protection Officer). All partners will also keep detailed documentation on data control and processing activities.

Informed consent for data sharing and long term preservation is included in questionnaires dealing with personal data.

Data sharing will be defined before data collection so that users are able to consent. If data is shared to unanticipated parties, this data will be anonymised or be evaluated if compatible grounds for scientific research apply.

Co-funded by the Horizon 2020
Framework Programme of the European Union

6 OTHER

- **Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any).**

Efforts are underway to implement data management principles throughout all partners.

Co-funded by the Horizon 2020
Framework Programme of the European Union