

Bloomen

Blockchain for Creative Work

Blockchains in the new era of participatory media experience

HORIZON 2020

*762091 – BLOOMEN - H2020-ICT-2016-2
ICT-19-2017 Media and content convergence*

D1.1 Initial Data Management Plan

Version:	1
Date:	30.03.2018
Authors:	WORLDLINE
Internal Reviewers:	Amaryllis Raouzaiou, ATC Quimi Luzón, BMAT

Worldline



Co-funded by the Horizon 2020 programme
of the European Union

Table of Contents

1	Introduction	3
2	Data Summary	4
3	FAIR datasets	6
3.1	Datasets identification and description	6
3.2	Data Management Platforms.....	7
3.3	FAIR data template	7
3.4	Source code	8
4	Other data management aspects	10
4.1	Allocation of resources.....	10
4.2	Data security	10
4.3	Ethical aspects.....	10
5	Conclusions	11

1 Introduction

Bloomen participates in the H2020 Open Research Data Pilot (ORD Pilot), as specified in section 2.2.1.6 on Part B of the DoA. The Data Management Plan (DMP) of Bloomen will provide an overview of the available research data arising from the project, the data accessibility, management and terms of use. The DMP will follow the template that the European Commission suggests in the “Guidelines on FAIR Data Management in Horizon 2020”, current version is 3.0, dated 26 July 2016, consisting of a set of questions that the project shall address and properly answer with a level of detail appropriate to the project. 'FAIR' data refers to data that is Findable, Accessible, Interoperable and Re-usable.

According to these guidelines, the DMP will include the following sections:

1. Data Summary
2. FAIR Data
 - a. Making data findable, including provisions for metadata
 - b. Making data openly accessible
 - c. Making data interoperable
 - d. Increase data reuse (through clarifying licences)
3. Allocation of resources
4. Data security
5. Ethical aspects
6. Other

This document is the initial version of the DMP, and it does not intend to answer these questions, but to present the information on how the actual DMP will be put together and its contents when data from the project will have become available. This document will be updated during the project and populated with the answers to the questions, concluding with the final DMP, deliverable D1.2, which is due in month M30.

2 Data Summary

According to DMP guidelines, this section will address the following questions during the project lifetime:

1. What is the purpose of the data collection/generation and its relation to the objectives of the project?
2. What types and formats of data will the project generate/collect?
3. Will you re-use any existing data and how?
4. What is the origin of the data?
5. What is the expected size of the data?
6. To whom might it be useful ('data utility')?

From all these questions, in this initial DMP we are starting to address the first two questions, while the other remaining four questions will be analysed as soon as the progress of the project provides more concrete information on the datasets.

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Essentially, data generated during the project will come from the specific needs of each of the Bloomen pilots, but also some data will be generated for measurement and assessment purposes of the Bloomen platform. This data generation processes are in fact directly connected to Bloomen project objectives:

- *Objective 1: The project will design and implement a new paradigm of a distributed multiplatform architecture for content creation, sharing and consumption based on blockchains.*
While the content itself will not be generated by the Bloomen platform, the management of those contents will finally derive into blockchain transactions containing associated Bloomen metadata. These metadata will later be relevant not only for pilots and Bloomen platform evaluation, but also for third parties after the end of the project with similar pilots and intending to adopt Bloomen solution.
- *Objective 2: The project will provide a set of innovative services in order to facilitate the convergence of the media delivery platform and support the new business models of the modern media industry.*
Bloomen will be providing several services, such as identity management, copyright management or micropayments, which will be generating additional metadata over the digital contents which, as mentioned before, will be relevant for evaluation and solution reuse purposes. Whether these datasets will be made publicly available or not, will have to be decided case by case depending on several sharing criteria such as their nature, ownership or exploitability. Preference will always be given to openness, while private datasets shall be the exception, properly justified.

- *Objective 3: The project will validate the new Bloomen offering through real life use cases in particular in the music industry, news media industry and video on demand industry.*

The execution of the real life use cases will require access to digital content and management databases, and will also generate new data and metadata. The availability of such datasets for public domain will be entirely dependent on each use case. If such datasets are already open they will continue being open, but those of private nature will not be disclosed unless the corresponding use case owner has the right to take such decision and decides to do so.

- *Objective 4: To provide a blueprint of best practices and disruptive business models on how blockchains can be effectively applied for transforming the media industry for the benefit of all actors in the value chain.*

Both best practices and business models will be made public, and whenever based on open datasets, such datasets will be also made available. However, as mentioned before, when referring to specific use cases, some related datasets may remain as private if that is the decision of its use case owner.

- *Objective 5: The project will maximize the impact of its results through dissemination, exploitation and community building activities.*

Even if these activities will not generate or manage any specific project dataset, in general all data related to stakeholders involved in community building will be made open, provided they do not include any private data, which will be either anonymized if possible or completely removed prior to disclosure.

What types and formats of data will the project generate/collect?

Even if the specific data formats will be further elaborated during the project, data generated by the Bloomen platform will mostly consist of blockchain transactions that may also include edited metadata, but also other measurement and traffic data from the peers of the Bloomen blockchain network. In addition to the data generated by Bloomen itself, the execution of the Bloomen pilots will also require accessing and collecting different types of data related to the digital contents being managed by the Bloomen platform and services.

As soon as each specific dataset is identified, the project will decide on the specific format considering that, as explicitly mentioned in the DoA, the overall goal is to as much as possible, use not only open formats to store the data but also open source software to provide the scripts and other metadata necessary to reuse it.

3 FAIR datasets

Bloomen will validate and demonstrate the innovations designed and implemented within the project through three pilot use cases, as defined in deliverable “D2.1 Use Case Descriptions and KPIs”. These pilot use cases will include several data related activities, such as creation, collection, storage, management, processing or deletion of datasets, which in general means that while some datasets will be used as an input to pilot activities, some other datasets will be generated during piloting activities.

Regarding the datasets being used as an input to pilot activities, the project will consider which of these data can be made available for open access and the applicable procedures if any personal data is involved, as specified in deliverable “D7.1 Protection of Personal Data (POPD) - Requirement No.1”. Therefore, the identification of datasets will take place before starting the pilot operation phase, followed by the acquisition of the informed consent and the ethical approvals when required.

Regarding the datasets generated by the pilots, they will also be considered for open access. However, at this point it is too early to identify such datasets, this will happen once the Bloomen platform has been implemented and the pilots have been completely designed. It is expected that most datasets, if not all, related to Bloomen platform evaluation will be made available for open access, after following the same procedures for the input datasets described above.

3.1 Datasets identification and description

As specified in the guidelines of the European Commission on Data Management, the data to be made available for open access will have to be described using the following dataset description template. These descriptions will be stored in the project’s internal repository and will be provided in the next update of this deliverable (D1.2 – Final Data Management Plan).

Dataset reference and name	Identifier for the dataset to be produced.
Dataset description	Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.
Standards and metadata	Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.
Data sharing	Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other

	tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).
Archiving and preservation	Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

3.2 Data Management Platforms

Bloomen will use OpenAIRE (www.openaire.eu), in cooperation with re3data (www.re3data.org), to select the proper open access repository and/or deposit publications for its research results storage, allowing also for easy linking with the project and facilitating open access to scientific publications. This will increase the accessibility to the obtained results by a wider community, which can be further enhanced by including the repository in registries of scientific repositories, such as DataCite (www.datacite.org), OpenDOAR (www.opendoar.org), or Zenodo (www.zenodo.org). These are the most popular registries for digital repositories and along with re3data, they are collaborating to provide open research data.

Additionally, Bloomen will take advantage of the built-in license wizard of the EUDAT B2SHARE tool (b2share.eudat.eu) to select the most adequate license for the project research data.

3.3 FAIR data template

'FAIR' data refers to data that is Findable, Accessible, Interoperable and Re-usable. During the project lifetime, and according to DMP guidelines, the following questions shall be addressed:

Making data findable, including provisions for metadata

- Outline the discoverability of data (metadata provision)
- 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?
- Outline naming conventions used
- Outline the approach towards search keyword

- Outline the approach for clear versioning
- Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how

Making data openly accessible

- Specify which data will be made openly available. If some data is kept closed provide rationale for doing so
- Specify how the data will be made available
- 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)?
- Specify where the data and associated metadata, documentation and code are deposited
- Specify how access will be provided in case there are any restrictions

Making data interoperable

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

Increase data re-use (through clarifying licences)

- Specify how the data will be licenced to permit the widest reuse 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
- 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
- Describe data quality assurance processes
- Specify the length of time for which the data will remain re-usable

This questions template will be filled as soon as datasets get defined as described in previous section 3.1, and will be provided in the next update of this deliverable (D1.2 – Final Data Management Plan).

3.4 Source code

In addition to the open data discussed above, Bloomen will also make available the generated software and its source code to the Open Source Community from the very beginning of the project. To this end, the entire source code will be available from the Bloomen's github account (github.com/bloomenio). The Open Source license of the source code is still under discussion, but a dual licensing scheme is being considered in order to protect the business exploitation perspectives of the partners. Dual licensing mixes several of the introduced generic software business models. Duality means that

both the free software distribution mechanism and traditional software product business are combined. There is technically only one core product but two licenses: one for free distribution and free use and another one for commercial use (proprietary). Since the specific strategy with regards to licensing is very related to the development of the Bloomen business models, the licensing model will be first outlined at deliverable “D6.6 Initial Exploitation Strategy and plan”.

Finally, all required documentation for installation instructions, developers’ guide, etc., will be provided in github and its respective wiki pages, which will be also referenced from the Bloomen web site.

4 Other data management aspects

The DMP guidelines also refer to the following aspects related to data management:

4.1 Allocation of resources

Since the very beginning of the design of the Bloomen project, data management was taken into consideration and, as leader of task T1.4 on Data Management, Worldline already allocated 2 PMs for this purpose. However, in addition to this specific effort, all use case partners and technical partners, with their related role, are involved in data management activities, either collecting, processing, or creating datasets, and the corresponding effort is embedded into the tasks in which they are undertaking these activities. Hence, all related costs for data management are already covered by the Bloomen project and no additional resources will be needed.

4.2 Data security

Any issue regarding the Protection of Personal Data has already been discussed in deliverable “D7.1 POPD - Requirement No.1”, and hence it is not repeated here. In addition to personal data protection, Bloomen will use state-of-the-art technologies for secure storage, delivery and access of personal information, as well as managing the rights of the users. In this way, there is complete guarantee that the accessed, delivered, stored and transmitted content will be managed by the right persons, with well-defined rights, at the right time.

State-of-the-art firewalls, network security, encryption and authentication will be used to protect collected data. Firewalls prevent the connection to open network ports, and exchange of data will be through consortium known ports, protected via IP filtering and password. Where possible (depending on the facilities of each partner) the data will be stored in a locked server, and all identification data will be stored separately.

A metadata framework will be used to identify the data types, owners and allowable use. This will be combined with a controlled access mechanism and in the case of wireless data transmission with efficient encoding and encryption mechanisms.

4.3 Ethical aspects

In addition to Protection of Personal Data, which has already been discussed in deliverable “D7.1 POPD - Requirement No.1”, Bloomen does not include any other ethical aspects that should have to be considered. Hence, all information related to ethical aspects can be found in that D7.1 deliverable.

5 Conclusions

This deliverable provides the initial Data Management Plan of Bloomen. It is in fact a placeholder for the different aspects that need to be tackled according to DMP guidelines and as soon as datasets get identified. The document defines how those datasets have to be properly described. While the project progresses, it will be identified what possible data or metadata can be opened to other parties, considering both datasets to be used as input to pilot activities, and datasets generated by the pilots themselves.

The document also discusses how and in which conditions source code will be made available. Finally, the document identifies the respective platforms that will host data and source code.

In this context, Bloomen will provide an updated and concrete Data Management Plan, including also the description of the identified datasets, on deliverable “D1.2 – Final Data Management Plan”.