Bloomen

Blockchains in the new era of participatory media experience

HORIZON 2020

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

D1.2 Final Data Management Plan

Version:	1
Date:	29/02/2020
Authors:	WORLDLINE
Type:	Report
Dissemination level:	Confidential

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

Table of Contents

List of abbreviations

CISAC	International Confederation of Societies of Authors and Composers
CMO	Collective Management Organisation
CWR	Common Works Registration
DDEX	Digital Data Exchange
DMP	Data Management Plan
EBU	European Broadcasting Union
EBU CCDM	EBU Class Conceptual Data Model
EXIF	Exchangeable image file format
FAIR	Findable, Accessible, Interoperable and Re-usable
GDPR	General Data Protection Regulation
ICPN	International Code Product Number
IPI	Interested Party Information
ISRC	International Standard Recording Code
ISWC	International Standard Musical Work Code
JSON	JavaScript Object Notation
KYC	Know Your Customer
MW	Musical Works
ORD Pilot	Open Research Data Pilot
SR	Sound Recordings

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 provides an overview of the available research data arising from the project, the data accessibility, management and terms of use. The DMP follows 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 has addressed and answered 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 includes 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

2 Data Summary

According to DMP guidelines, this section addresses 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')?

A first answer to the first three questions is provided in this section to highlight the general Bloomen approach, while more detailed answers to all questions are later provided in the sections of this document related to each Bloomen use case.

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

All data generated during the project come from the specific needs of each of the Bloomen pilots, while the Bloomen platform does not generate any data by itself. The data generation processes in Bloomen are in fact directly connected to the project objectives:

- *Objective 1: The project will design and implement a new paradigm of a distributed multi-platform architecture for content creation, sharing and consumption based on blockchains.*

  While the content itself is not generated by the Bloomen platform, the management of those contents finally derive into blockchain transactions containing associated Bloomen metadata. These metadata are later 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 provides several services, such as copyright management or tokenization, which are generating additional metadata over the digital contents which, as mentioned before, are relevant for evaluation and solution reuse purposes. Whether these datasets are made publicly available or not, has been decided case by case by each pilot, depending on criteria such as their nature, ownership or exploitability.

- *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 requires access to digital content and management databases, thus generating new data and metadata. The availability of such datasets for public domain, in total or partially, is entirely dependent on each use case. Datasets with a private nature are not disclosed unless all GDPR requirements, such as user consent, are met and the corresponding use case owner, who has the right to take such decision, 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 have been made public. However, as mentioned before, when referring to specific use cases, some datasets remain as private since 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.*

  These activities do not generate or manage any specific dataset suitable for publication.

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

The execution of the Bloomen pilots require the accessing and collecting of different types of data related to the digital contents being managed by the Bloomen platform and services. The specific data formats are detailed for each of the three pilots in sections 3.1, 3.2 and 3.3.

Will you re-use any existing data and how?

The potential re-use of existing data is entirely dependent on each pilot. While there is a very clear case for data re-use in the music pilot with catalogues and claims, which are central to the use case, only some secondary data sources may be re-used for the Photo and WebTV use cases.

3 FAIR datasets

Bloomen validates and demonstrates the innovations designed and implemented within the project through three pilot use cases. These pilot use cases include several data related activities, such as creation, collection, storage, management, processing or deletion of datasets, which in general means that while some datasets are used as an input to pilot activities, some other datasets are being generated during piloting activities.

3.1 Data in Music use case

3.1.1 Data summary

For all the data managed under this use case, the following questions need to be answered:

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

The data itself is the purpose of the project. In the music use case, we want to explore if the users can create and consume better music rights data with the tools the project provides.

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

The project will interact mainly with data that models two kinds of musical assets: musical works and sound recordings, represented in JSON format, represented by the core metadata that describes the assets. On the other side, the rights metadata describing right holder information is generated by the users of the pilot, which is stored in a blockchain.

3. Will you re-use any existing data and how?

Yes. The software developed in the project offers import tools which make it possible to re-use datasets: the Bloomen Management Portal allows the import of catalogs of musical assets, and the Bloomen Decentralized Rights Management app allows the import of right holder claims, which can be borrowed from other platforms.

4. What is the origin of the data?

The metadata of musical assets can be taken from public or private repositories provided by the pilot testers. The right holder data will be provided by the users of the platform.

5. What is the expected size of the data?

The amount of data managed by the platform depends on the nature of the tests. For scalability tests, BMAT counts with datasets of several million of assets that can be loaded into the platform thanks to the data import functionalities of the software.

6. To whom might it be useful ('data utility')?
The data collected and generated by the system will be useful for interested parties on the musical assets represented in there, mainly right holders, as well as the stakeholders that make use of it as part of their workflows, like right management organizations.

3.1.2 Data models and Identified datasets

The deliverable D1.1 "Initial Data Management Plan" included a preliminary data model for the music use case defined during the requirements phase. Throughout the development of the project, the data model has been refined to meet the needs of the use case:

Music Data Model	
CMO	This entity represents the Collective Management Organisation, which groups the members they represent
Member	This entity represents the member affiliated to a given CMO: Name CMO ID IPI Name Number Country Creation Date
User	This entity represents the user of the system that acts on behalf of the member affiliated to a given CMO: System User ID First Name Last Name Member Role (SuperAdmin, Admin or User) Creation date
Musical Asset	The entity that represents Musical Works (MW) or their Sound Recordings (SR). They are defined by a set of core metadata (international identifiers, title, contributors, etc), and they have rights holders attached to them.

Metadata	**Musical Work (MW):** **Core:** ISWC Code Original Title Alternative Titles Creators (Name, IPI Name Number, Role) **Other:** Associated Performers Associated ISRCs **Sound Recordings (SR):** **Core:** ISRC Code Main Artist Featured Artists Title Version Title Duration Year of Recording Territory of Recording Language of Performance Original Release Date Original Release Label Creators isVideo **Other:** Release Information (Title, Artist, ICPN, Number of Tracks, Label, Duration, isCompilation)
Rights Claim	The entity that represents the information of a rights holder's claim over a musical asset: Musical Asset ID (ISRC or ISWC) System User ID Right Holder Name Right Holder Proprietary Asset ID Right Holder Role Right Types Territories Start Date End Date Split Status (Claimed/Conflict)

From this data model, the music use case has identified and detailed the following datasets:

Dataset reference and name	Bloomen Music Dataset

Dataset description	The dataset will contain 2 types: musical work and sound recording, containing core metadata to describe the asset and right holder information.
	It will be represented in JSON format.
	The user is in charge of generating the data, and it is useful for the right holder of the asset as well as for the organisation appointed to manage the royalties generated by the assets in the dataset.
	It has the following structure:
	
	"musicalWork": {
	"ISWC": <text> (ISWC format),
	"originalTitle": <text>,
	"creators": [{
	"name": <text>,
	"IPINameNumber": <text> (IPI format),
	"role": <enum>
	}],
	"alternativeTitles": [<text>],
	"associatedPerformers": [<text>],
	"associatedISRCs": [<text> (ISRC format)],
	"rights": [{
	"rightsHolder": {
	"name": <text>,
	"IPINameNumber": <text> (IPI format),
	"role": <enum>
	},
	"rightsHolderProprietaryId": <text>,
	"territories": [<enum>],
	"startDate": <date>,
	"endDate": <date>,
	"mechanical": {
	"affiliationSociety": <text>,
	"ownershipSplit": <float> (percentage),
	"collectionSplit": <float> (percentage)
	},
	"performance": {
	"affiliationSociety": <text>,
	"ownershipSplit": <float> (percentage),
	"collectionSplit": <float> (percentage)
	},
	"synchronisation": {
	"affiliationSociety": <text>,
	"ownershipSplit": <float> (percentage),

	"collectionSplit": <float> (percentage)
	}
	}]
	}
	
	musicalWork.creators.role:
	"Adapter"
	"Arranger"
	"Lyricist"
	"Composer"
	"ComposerLyricist"
	"SubArranger"
	"SubAuthor"
	"Translator"
	"IncomeParticipant"
	
	musicalWork.rights.rightsHolder.role:
	"Writer"
	"OriginalPublisher"
	"SubPublisher"
	"RoyaltyAdministrator"
	
	musicalWork.rights.territories:
	alpha-2/ISO 3166-1 codes
	
	"soundRecording": {
	"ISRC": <text> (ISRC format),
	"mainArtist": <text>,
	"featuredArtists":[<text>],
	"title": <text>,
	"versionTitle": <text>,
	"duration": <int>,
	"yearOfRecording": <int>,
	"territoryOfRecording": <enum>,
	"languageOfPerformance": <enum>,
	"originalReleaseDate": <date>,
	"originalReleaseLabel": <text>,
	"creators": [<text>],
	"isVideo": <bool>,
	"releases": [{
	"title": <text>,
	"artist": <text>,
	"ICPN": <text> (ICPN format),
	"numberOfTracks": <int>,
	"label": <text>,
	"duration": <int>,

	"isCompilation": <bool>
	}],
	"rights": [{
	"rightsHolder": <text>,
	"rightsHolderProprietaryId": <text>,
	"territories": [<enum>],
	"startDate": <date>,
	"endDate": <date>,
	"split": <float> (percentage),
	"useTypes": [<enum>],
	"CMO": <text>
	}]
	}
	
	soundRecording.rights.useTypes:
	"PublicPerformance"
	"Airlines"
	"RadioBroadcasting"
	"RadioDubbing"
	"TVBroadcasting"
	"TVDubbing"
	"BackgroundMusic"
	"BackgroundMusicDubbing"
	"KaraokePublicPerformance"
	"KaraokeDubbing"
	"KaraokeOnDemand"
	"KaraokeOnDemandDubbing"
	"CableRetransmission"
	"RadioSimulcast"
	"Webcast"
	"TVSimulcast"
	"CatchUpTV"
	"PrivateCopying"
	"RingbackTones"
	
	soundRecording.rights.territories:
	alpha-2/ISO 3166-1 codes
	
	soundRecording.languageOfPerformance:
	alpha-3/ISO 639-2 codes

Example of a musical work asset:

{
"ISWC": "T9204649558",
"originalTitle": "SHAPE OF YOU",
"creators": [{
"name": "SHEERAN ED",
"IPINameNumber": "00583552527",
"role": "ComposerLyricist"
},
{
"name": "MCDAID JOHN",
"IPINameNumber": "00412720203",
"role": "ComposerLyricist"
},
{
"name": "MAC STEVE",
"IPINameNumber": "00257395141",
"role": "ComposerLyricist"
},
{
"name": "COTTLE TAMEKA D",
"IPINameNumber": "00338239158",
"role": "ComposerLyricist"
},
{
"name": "BURRUSS KANDI L",
"IPINameNumber": "00338170958",
"role": "ComposerLyricist"
},
{
"name": "BRIGGS KEVIN",
"IPINameNumber": "00344353278",
"role": "ComposerLyricist"
}],
"alternativeTitles": ["SHAPE OF YOU (STORMZY REMIX)", "SHAPE OF YOU [OFFICIAL LYRIC VIDEO]", "SHAPE OF YOU (LIVE FROM THE 59TH GRAMMYS)", "SHAPE OF YOU - ACOUSTIC"],
"associatedPerformers": ["WALK OFF THE EARTH", "WALE FEATURING ED SHEERAN", "SING2PIANO", "SHEERAN ED", "MADILYN BAILEY", "GALANTIS", "FAME ON FIRE", "ED SHEERAN", "BOYCE AVENUE"],
"associatedISRCs": ["GBAHS1600463", "GBAHS1700245", "GBAHS1700200", "GBAHS1700196", "GBAHS1700651"]
}

	Example of a sound recording asset:
	
	{"ISRC": "BER181131702",
	"mainArtist": "EX TEMPORE",
	"featuredArtists": ["FLORIAN HEYERICK"],
	"title": "DURCH DIE HERZLICHE BARMHERZIGK",
	"versionTitle": null,
	"duration": 227,
	"yearOfRecording": 2002,
	"territoryOfRecording": "BE",
	"languageOfPerformance": null,
	"originalReleaseDate": "2002-01-01",
	"originalReleaseLabel": "OUTHERE SA",
	"creators": ["GOLDBERG JOHANN GOTTLIEB"],
	"isVideo": false,
	"releases": [{
	"title": "GOLDBERG JL BACH & KREBS",
	"artist": null,
	"ICPN": "5400439003170",
	"numberOfTracks": null,
	"label": "RICERCAR",
	"duration": null,
	"isCompilation": false
	}

Standards and metadata	There are 2 main standards that deal with this kind of information: CISAC's CWR for musical works and DDEX MLC for sound recordings. These standards define not only the data structure but also the protocols for communicating this kind of information. They are very complete but complex in excess for a first proof of concept of the system. They could be adopted as the standard way to import and export data to and from the system as following steps.
Data sharing	The data entered into the system and generated by the users is shared among participants in the platform, applying some privacy rules depending on the role of the user (e.g. restricted access to musical rights claiming data). There are tools developed in the project to enable import and export of data in JSON and CSV format. The repository of musical repertoire is stored in the platform servers. The part of the dataset that contains right holders information is stored in the blockchain and should not be shared without prior consent of the right holders.

Archiving and preservation	The data will be preserved during the project lifetime. Due to the nature of the system, there will be copies of the repository on several nodes. For the expected volume of data, and considering that the dataset contains only text, the associated costs for preserving the data are negligible.

Apart from the data generated during the tests of the tool developed throughout the project, BMAT has identified a compatible dataset that can be imported into the system. BMAT counts with DIG-IT, a rights management platform to deal with the copyright management of sound recordings consumed in Italy. This dataset contains more than 11 million sound recordings, described with a limited subset of fields, combined with a dataset of more than 4 million related right holder claims that can be used to test the music use case. This dataset belongs to BMAT and the customers of DIG-IT, but it can be used for the pilot test of Bloomen. Although the dataset was designed to work with DIG-IT, it is possible to reuse it in the Bloomen platform thanks to the importing tools of the Bloomen software, which ensure its interoperability.

3.1.3 FAIR data template

'FAIR' data refers to data that is Findable, Accessible, Interoperable and Re-usable.

Making data findable, including provisions for metadata

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

  Both the Bloomen management platform and the Decentralized Rights Management app have full text search capabilities which allow the retrieval of the information based on the metadata of the assets.

- *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?*

  The assets are identified using the industry standard codes: ISWC for musical works and ISRC for sound recordings.

- *Outline naming conventions used*

  Naming conventions, including the vocabulary of enumerable objects, are taken from DDEX, the music data exchange industry standard.

- *Outline the approach towards search keyword*

  All relevant metadata fields are indexed

- *Outline the approach for clear versioning*

  The Bloomen management platform stores every change on the metadata of the assets stored. On the other hand, changes in right holder data are traceable because this kind of data is stored in a blockchain.

- *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*

  The metadata created is based on the CWR and DDEX standards, music industry references.

Making data openly accessible

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

  Musical repertoire data can be made openly available, but it lacks special interest, since it can be found in other public repositories. The right holder data cannot be made publicly available out of the Bloomen platform without the consent of the right holders, since this data belongs to them.

- *Specify how the data will be made available*

  The musical repertoire data can be made available through the Kendraio dashboard.

- *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)?*

  The data used or generated by the project can be accessed through the Kendraio software and the Decentralized Rights Management app, both developed within the scope of the project, and described in related deliverables.

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

  The data is deposited in 2 places: the MongoDB database for the musical repertoire, and in a blockchain for the right holder data. The documentation and code is part of the deliverables of the project.

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

  The access will be granted to pilot testers by creating a user profile for them with a given role and membership which defines the scope of the data they can access.

Making data interoperable

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

  The data is interoperable because the metadata vocabularies are based on music industry standards (CWR and DDEX). Besides that, the Kendraio app offers connectors to load data from other systems.

- *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?*

  The vocabulary used for all data types is taken from CWR and DDEX standards. The identifiers adopted for the assets (ISWC and ISRC) are also industry standards.

Increase data re-use (through clarifying licences)

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

  The data coming from BMAT systems can be reused for Bloomen pilot tests. The re-usability of the right holder information generated throughout the project needs to be agreed with the pilot testers.

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

  No data embargo is needed. The data can be made available for re-use whenever necessary if there is an agreement with the right owners.

- *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 can be re-used by third parties if right holders agree to share it, by giving the third party access to the Bloomen platform.

- *Describe data quality assurance processes*

  The data quality assurance itself is the purpose of the project. In the music use case, we want to explore if the users can create and consume better music rights data with the tools the project provides.

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

  The conditions of how the data can be reused, including the length of time, is determined by the different right holders.

3.2 Data in Photo use case

3.2.1 Data summary

The purpose of data collection for the Bloomen Photo use case spans several key aspects, all relevant for the photo platform: Identification, verification, trustable rights management and simplification of workflows. These are the key reasons for collecting data. The type of data collected will be:

- personal and institutional with a scope on reliable identification (Who is the owner/creator? Who is the user? KYC-relevant);
- item-based, as to determine and verify that a photo (item) has been created, is uploaded, has metadata and is owned by an identifiable photographer or an agency;
- attribution and licenses, in order to identify the owner of a work (person or company), the key elements of the license (exclusive, non-exclusive) and the usage rights agreed upon (non-exclusive, exclusive), the usage rights over time (usage period), regional usage rights (geography), other rights (popularity or success of a picture over time)
- personal data combined with item-based data help to determine a trustable level of rights, ownership and usage rules;
- finally, based on deploying blockchain technologies, the application has a higher chance of exploitation when in effect the complexity of managing people and assets in high volume are simplified, e.g. through the use of rules-based systems or “smart contracts”.

Re-use: Some of the data can be re-used from existing sources. For example, should there be a third party service for reliable identification which is trustable, we could agree to anonymity (e.g. not naming the author/creator of a photo). Though it has to be kept in mind that in the media industry, authors and creators usually want to be associated with their work, with very few exceptions (e.g. whistleblowing, unstable political situations).

Other data will have to be generated, e.g. a reliable system to identify and collect data about photos, ranging from available EXIF data to creation data/ownership and potentially other data point adding to the searchability of photos (e.g. Artificial Intelligence applications “looking” at photos and providing information about persons, places, angle, color, size of the photo, resolution, geographic location, etc.). We expect such data to be available sometimes, but often in an inconsistent status, which poses a problem.

Origin of the data: The data will be either auto-generated, either from other systems, databases or cameras when taken (e.g. mobile phones) and in other cases will be sourced from the creator/photographer or the media company using the photos (when? where? How long? How often? etc.).

Current technical storage model for data (demonstrator/development). Bloomen aims to store only minimal data of users, for the single purpose of processing uploads, licenses and payments.

- Identity of a user, verified for further transactions
- Photos or any other assets which are attributed to the user
- Photos are stored on Amazon S3, which has procedures for GDPR-compliance
- Data of users is stored in MongoDB, the database technology has procedures and modules to achieve GDPR-compliance
- Special protection and procedures would be needed to shelter long-term user profiles, e.g. how many photos have been sold by a photographer and to whom, how many photos or other media assets a media organization has bought over time.

Size of the data: At this moment it is not possible to reliably determine the size of the data. The expectation is that a successful, deployed system will handle Gigabytes of photo data. In principle there will be two data collections, one with primarily metadata and blockchain generated hashes, another with the items itself. The media assets themselves (photos, videos, etc.) are not not stored on the blockchain, instead a hash or other means of identification of a specific media asset is stored, but nothing more. Given performance issues the general direction as of early 2019 is to separate the photo files from the blockchain-based identification and metadata. In effect, the data can grow to large sizes, given the number of photos taken and used in a typical media organization. At the same time these storage needs are already covered and can be handled by existing systems, on premise or in a cloud.

Data utility: It should be understood that the purpose of the data collection and the usefulness of the data is to make it beneficial for both sides - the creator/photographer and the media organisation. Only when this is achieved there is a good chance of commercial exploitation or other lasting ways to further develop the approach of Bloomen photo. Further, very different to an advertising-based/”free” offering the data is not collected without, but with the consent of all participants.

Instead long-term data profiles of Bloomen Photo users would be based on a contract and data processor agreement which would define the specific purpose of the relationship. In order to fulfil the purpose of Bloomen Photo the platform must keep minimal data for both photo editors and photographers, to identify, maintain and notify these users. Further, minimal profile information is needed for financial transactions on behalf of the aforementioned users: Should e.g. a photographer sell a photo he/she is entitled to a payment, be it in a currency or a crypto coin. Same for the photo editor who acquired one or multiple photos, in terms of having made the payment, having acquired the rights to use a picture.

In the case of further development of Bloomen Photo towards a commercial platform we would need to establish a legally sound, GDPR-compliant role as either a

contractor (and securitization of the rights of the user) or a data processor. The term data processor would apply only to an extent, as the data would usually not leave the system for any other purposes.

High quality metadata: For a reliable and useful platform, particularly for exploitation and towards commercial sustainability high quality metadata will be very important. The key approach is to simplify the identification of a photo, its license and its owner, based on standards accepted and maintained by public bodies such as the EBU (European Broadcasting Union).

Use of standards: There are two specific metadata standards in use in the media industry: EBU Core[1] and - to some extended level - EBU CCDM[2]. The EBU (European Broadcasting Union) is the world's largest organization of public service media organisations and therefore a standard supported by the EBU is the best way forward for standardized data management and ontologies.

For clarification: While EBU Cored defines the basics, EBU CCDM is an extended set of metadata classifications used for material in production in various phases.

> EBU Core is the flagship ontology maintained for usage by public media companies (and others in the media field). It is based on the well-established Dublin Core. Definition: EBUCore is a set of descriptive and technical metadata based on the Dublin Core and adapted to media. EBUCore is the flagship metadata specification of EBU,, the largest professional association of broadcasters around the world. It is developed and maintained by EBU's Technical Department. EBU has a long history in the definition of metadata solutions for broadcasters.[3]
>
> About EBU CCDM: The EBU Class Conceptual Data Model (CCDM) is an ontology defining a basic set of Classes and properties as a common vocabulary to describe programmes in their different phases of creation from commissioning to delivery. CCDM is a common framework and users are invited to further enrich the model with Classes and properties fitting more specifically their needs.[4]

Beyond this more general approach Bloomen Photo has identified a number of more specific data aspects, listed below.

3.2.2 Data models and Identified datasets

In D.2.2 “Bloomen Requirement Analysis”, the basic data models to describe the users, the assets and the transactions were already described. They reflect an early status, most of which has already become functional in the current Bloomen Photo demonstrator.

[1] EBUCore - https://tech.ebu.ch/files/live/sites/tech/files/shared/tech/tech3293.pdf
[2] EBU CCDM - https://tech.ebu.ch/docs/tech/tech3351.pdf
[3] Source: https://en.wikipedia.org/wiki/Metadata_standard
[4] Source: https://en.wikipedia.org/wiki/Metadata_standard

Preliminary Photo Data Model	
Users	Distinguish two roles: Publisher or photographer Authorisation data (username, password, etc) Name & Address Settings (privacy, payments, etc) Role (Consumer, Contributor) Reputation (reputation of the user, can be applied to both creators and media organisations over time)
Assets	Url (public file url) Type of asset: UGC, photo, special photo Rights (list of users that have rights using this file) Owner (the owner of the file) Date/time added Price (price to pay for publishing rights) Usage rights time (how long?) Usage rights region (where in the world?) Analytics (number of views, likes, etc) Keywords Description Geo-coordinates Hash for organisation
Transaction	From To Date Amount

At this point in time, work in the Photo use case has been more focused in the functional aspects and user interface, while the detailed definition of datasets is still in progress.

What is clear is that the project anticipates that a practical data plan will be important for the platform, should it be exploited commercially.

3.2.3 FAIR data template

'FAIR' data refers to data that is Findable, Accessible, Interoperable and Re-usable.

Making data findable, including provisions for metadata

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

  A key part of Bloomen Photo is to ensure that metadata identifying a specific photo (or other media asset such a video or a data visualization) is identifiable, even with years between recording and searching.

- *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?*

  Yes, though there is no final decision which identifiers would be used in production.

- *Outline naming conventions used*

  Given the status of the research project and current iterations such naming conventions are subject to change and not finalized at this moment.

- *Outline the approach towards search keyword*

  Bloomen Photo anticipates that not all submissions will come with the needed level of keywords, which is why there is an option for both roles (photographers as well as publishers) to add keywords to an asset.

- *Outline the approach for clear versioning*

  Will be handled through the blockchain hash, in order to enable separation of photos from a series or similar production - e.g. when the time of creation, author/creator, time of upload are almost similar.

- *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*

  Photos allow for a variety of metadata. Bloomen Photo has experimental features showing how to store Exif-Data from a camera for each photo while uploading. Further, the standards for metadata creation primarily depend on what a given media organization sets as expected general or internal standards.

Making data openly accessible

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

  The platform will make only selective data public. This would include data about the volume of usage, the number of people using the platform and the number of uploads, as well as other activity. For all this data to be published it will be mandatory that the data would not allow to identify individual users of the system or enable tracking down such users, even unintentionally. Data on the creator and any statistics associated with the creator as well as data about a publisher and any statistics associated with an account will be kept strictly private. Bloomen Photo might develop a system to capture general data on load and activity, but in a way not allowing direct identification of any user to the outside. Internally, based on contracts which need to be developed yet, there will be a profile of a user which is private, but can be accessed by the platform operators in disputes of rights, ownership, payments and licenses.

- *Specify how the data will be made available*

  The data will be made available through GitHub or linked to from GitHub should the stats be stored elsewhere.

- *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)?*

  The data which is made available will be accessible in common formats, primarily as .csv (Comma separated values)

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

  The data will be stored on the platform used in production, most likely a cloud platform where data loss risk is reduced to a minimum.

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

  For restricted parts of stored data such as users, transactions, etc. there will be no outside sharing as a rule. Exceptions could be made in cases of license disputes, based on European and German law.

Making data interoperable

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

  The data will be based on common and widely used standards, to enable easy usage of the data. For this purpose the project will be based on EBU Core, as mentioned above.

- *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?*

  The final system will be based on ontologies (EBU Core) used by the media industry or more specifically photo community.

Increase data re-use (through clarifying licences)

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

  The data will be kept in the project, as the statistical value to other projects or platforms is highly limited.

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

  By the nature of a blockchain-enabled photo platform with commercial exchange modules re-use of the core data does not have research interest and would not be in the interest of the participants.

- *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*

  A key point is full or at least adequate privacy for the users, mainly in terms of security for the transactions and therefore the business value of Bloomen Photo for intended participants.

- *Describe data quality assurance processes*

  The standards of data quality assurance must be set very high, although benefitting from intended usage of the platform, e.g. a user wants to be associated securely to his/her photography works. Same or similar motives apply to the publisher related users.

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

  As said, re-usability is not a key goal. Long-term data storage though is highly important. In later stages a long-term data storage plan will be vital for Bloomen Photo, in order to guarantee a quick way to identify who is a licensed user of a photo, be it on the creator side or publisher side.

3.3 Data in WebTV use case

3.3.1 Data summary

For all the data managed under this use case, the following questions need to be answered:

1. What is the purpose of the data collection/generation and its relation to the objectives of the project?
The purpose of the data collection/generation is to explore whether users' engagement will be increased through the WebTV use case via the tools created by the technical partners. The usage of the mobile wallet as a complementary tool to the WebTV platform will be key in assessing the value the technology generates.

2. What types and formats of data will the project generate/collect?
The project will interact with various kinds of data:

- Form registrations
- Video consumption analytics
- Purchases of virtual currency and blockchain transactions.

3. Will you re-use any existing data and how?
We will use existing data regarding video analytics to compare engagement from users.

4. What is the origin of the data?
The origin of the data is the users of the WebTV Platform and the mobile wallet.

5. What is the expected size of the data?
Especially for the 2nd iteration of the WebTV pilot, it is expected that data will be generated from up to a hundred users of the WebTV Platform and the mobile wallet.

6. To whom might it be useful ('data utility')?
There are various stakeholders who might find the data useful. From content creators, to streaming platforms, as well as authorities researching the area of blockchain technology. A key approach is to use the data strictly for purposes benefitting the users (content providers, WebTV platforms) and advancing the business utility for the users, not the platform in any way beyond enabling the provision of an environment to transact audiovisual content.

3.3.2 Data models and Identified datasets

In deliverable D2.2 “Bloomen Requirements Analysis”, several data models for the WebTV use case were already identified, which are valid going forward with research and development.

Preliminary WebTV Data Model	
Users	Contains all necessary information of the user of the system. ● Authorization data [browser identity/cookies] ● Role (can be a consumer,a copyright owner or both) ● Wallet data and public addresses
Assets	An entity that represents copyrighted content available for commercialization, distribution and access. **Video Content:** ● Video Title ● Year of production ● Production company name ● File location (URL) ● Tags for indexing ● Video Analytics (Views, Likes, etc) ● Hash
Rights Holder	An entity that represents the information of a rights holder for an asset. ● Name ● Contact Information ● Role ● Rights Type ● Start Date ● End Date

Tokenization	Entities which represent store of value, means of reimbursements as well as cryptographic delivery of the content. ● Virtual Currency ● Transaction Info (Source, Destination, Amount, Timestamp, Transaction Hash) ● Video Server Delivery Access Control

3.3.3 FAIR data template

'FAIR' data refers to data that is Findable, Accessible, Interoperable and Re-usable.

Making data findable, including provisions for metadata

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

  It is important for the metadata of assets transacted and the transaction details to be identifiable through the mechanisms built within Bloomen and the integration tools for third party WebTV providers.

- *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?*

  The data is identifiable through the parameters set within Bloomen in collaboration with the third-parties that integrate the Bloomen tools within their operations. This regards information collected about the users' browsers as well as the public keys used to transact copyrighted content and the information about this content which is stored on the blockchain.

- *Outline naming conventions used*

  At the current point in time, naming conventions are not finalized or set to be used as standardized methods.

- *Outline the approach towards search keyword*

  Search keywords range from the official titles of the audiovisual material transacted and their release information such as season number etc.

- *Outline the approach for clear versioning*

  The uploading of information and hashes on the blockchain is handled by a mobile wallet which can be updated to deprecate/alter the structure of the data.

- *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*

The metadata created in the WebTV use case is currently subject for consideration within a potential WebTV consortium that will be formed to agree on its structure. Currently the platform is integrated on just one WebTV provider without discussions or disclosure or metadata handling/structure.

Making data openly accessible

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

  In terms of the WebTV use case, the data that will be made public is in regard to the usage (downloads) of the mobile wallet, the structure and number of transactions and smart contracts. Data regarding the integration of the Bloomen tools to the WebTV platform of ANTENNA which has to do with the visits on the dedicated platform will also be made public. Additional data regarding the identification of the users that access the tools through cookies, which range from browser type to timestamps, are irrelevant to the project and will be not utilized in the future for any purposes.

- *Specify how the data will be made available*

  The GitHub repository of Bloomen will include references to where the data will be made available when it is finalized.

- *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)?*

  The data which is made available will be accessible in common formats, primarily as .csv (Comma separated values)

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

  The data will be stored on the platform used in production, most likely a cloud platform where data loss risk is reduced to a minimum.

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

  Data that is non-public, i.e the access information (cookies) of the users of the platform where the Bloomen is integrated, is collected through a private database of the provider and as a rule, there will be no outside sharing unless there is a legal purpose to do so.

Making data interoperable

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

  Data on the integrated solution of Bloomen WebTV is defined and described according to the html5 vocabulary for XHTML.

- *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?*

  Standard vocabulary for all datatypes will be used for the integrated solutions of Bloomen.

Increase data re-use (through clarifying licences)

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

  The data will be kept in the project, as the statistical value to other projects or platforms is highly limited.

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

  There is no intent for the data to be made available for re-use as the structure of the use case is tailored for the needs of a specific WebTV provider.

- *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 analytics regarding the usage of the Bloomen Wallet and the WebTV of ANTENNA will be useful for interested parties wanting to integrate the solution or create something similar in the future. The data for the users and the content themselves is not something that might be considered useful for outside parties.

- *Describe data quality assurance processes*

  In regard to the usage of copyrighted content, data quality assurance is of most importance as the licensors have strict requirements for the usage of these copyrights.

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

  Since re-usability is not a key goal of the use case, although the plan is to keep the data for the long-term, no planning regarding its reusability is fixed.

3.3 Data in the Bloomen platform

The Bloomen Platform acts as a wrapper for the workflows of the pilots, therefore it does not store any more data than is needed for each of the pilots to function. It provides an interface to interact with the saved data, whether it is from the database, the cloud file storage or the blockchain.

Music use case data workflow

As far as the Music use case is concerned and the smart contracts deployed, in order to facilitate the functionalities needed for the music asset rights conflict identification and resolution, only few non sensitive information items are stored in the blockchain. More specifically, in order to uniquely identify the music assets two standards are used: ISRC (International Standard Recording Code) and ISWC (International Standard Musical Work Code). Only these types of unique identifiers are stored in the blockchain. These unique identifiers are publicly available and are not considered sensitive information. Other generic data concerning the sound recordings and music saved in the platform are saved in the database.

Photo use case data workflow

The Photo use case app uses the platform to save regular data, such as user information, most photo metadata, etc to the database. The photos themselves are uploaded to an amazon s3 file storage and retrieved only through the Bloomen platform for users that have access to them, either the creators or users that have purchased rights to use them through the Bloomen blockchain. When a photo is uploaded a transaction with the blockchain is created and a unique identifier for the photo is saved in the blockchain along with the price of the asset. Lastly when an asset is purchased the transaction is also saved in the blockchain and the user or organisation purchasing it is added to the ones that have rights on the asset.

WebTV data workflow

The WebTV use case only stores purchase data and access records to videos made from an anonymous identity that cannot be attributed to a specific end user. This data is stored within the Smart Contracts deployed in Alastria and on the end of the WebTV platform, only browser data (cookies) are stored to verify whether users have access copyrighted content.

4 Other data management aspects

4.1 Data Management Platforms

Bloomen has considered various open data platforms such as 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 was motivated by the wish to 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). Moreover, the project has produced scientific results which have been published under green and gold open access schemes.

4.2 Source code

In addition to the open data discussed above, Bloomen also makes available the generated software and its source code to the Open Source Community. To this end, the entire source code is available from the Bloomen github account (github.com/bloomenio). Regarding licensing, the Bloomen consortium decided that the Open Source license of the source code for all Bloomen components is MIT license, and hence all Bloomen source code is available in GitHub under that license.

Finally, all required documentation for installation instructions, developers' guide, etc., is also provided in GitHub and its respective wiki pages, which are also referenced from the Bloomen web site.

4.3 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, have been 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 have been required.

4.4 Data security

Any issue regarding the Protection of Personal Data was already discussed in deliverable “D7.1 POPD - Requirement No.1”. In addition to personal data protection, Bloomen uses 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 is managed by the right persons, with well-defined rights, at the right time.

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

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

4.5 Ethical aspects

In addition to Protection of Personal Data, which was already 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 final Data Management Plan of Bloomen, a summary for the different aspects that have been tackled according to DMP guidelines related to Bloomen identified datasets. The document describes these datasets, which data or metadata created, re-used or managed by the three different pilots can be shared with other parties, and in which way they can be accessed.

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