Horizon 2020

fashion BRAIN project

Understanding Europe's Fashion Data Universe

Data Management Plan

Deliverable number: D8.2

Version 4.0

Funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 732328

Project Acronym:	FashionBrain
Project Full Title:	Understanding Europe's Fashion Data Universe
Call:	H2020-ICT-2016-1
Topic:	ICT-14-2016-2017, Big Data PPP: Cross-sectorial and cross-lingual data integration and experimentation
Project URL:	https://fashionbrain-project.eu

Deliverable type	Report (R)
Dissemination level	Public (PU)
Contractual Delivery	30 June 2017
Resubmission Delivery Date	08 February 2019
Number of pages	35, the last one being no. 27
Authors	Alessandro Checco, Jennifer Dick - USFD
Peer review	Jen Smith - USFD (Research Data Management/The University Library)

Change Log

Version	Date	Status	Partner	Remarks
1.0	12/01/2018	Final	USFD	Rejected 15/03/2018
2.0	20/04/2018	Resubmitted Final	USFD	Rejected 15/10/2018
2.1	25/01/2019	Draft	USFD	
3.0	04/02/2019	Resubmitted Final	USFD	Rejected 06/02/2019
4.0	08/02/2019	Resubmitted Final	USFD	

Deliverable Description

Data Management Plan as required by the participation in the Open Research Data Pilot in Horizon 2020. This document contains guidelines and information from the project partners on how data involved in the project is stored and processed.

Abstract

This document describes the Data Management Plan (DMP) of the FashionBrain project. It is a short plan that outlines what data will be generated or collected, how data will be managed, the standards in use, the workflow to make the data accessible for use, reuse and verification, and which plans for data sharing and preservation exist ensuring that data are well-managed.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

BigComp 2019	6th IEEE International Conference on Big Data and Smart Computing
CC-BY	Creative Commons Attribution licence
CCG	Combinatory Categorial Grammar
CD	Centroid Decomposition
CLEF	Cross-Language Evaluation Forum
COLING 2018	27th International Conference on Computational Linguistics
CoNLL	Conference on Natural Language Learning
CQE	Continuous Query Engine
CSV	Comma Separated Values
DMP	Data Management Plan
EUDAT	The European Data Infrastructure
FaBIAM	FashionBrain Integrated Architecture
FAIR	Findable, Accessible, Interoperable and Reusable
GB	Gigabyte
ICDE 2019	35th IEEE International Conference on Data Engineering
ID	Identifier
IDEL	In-Database Entity Linking
IoT	Internet of Things
IP	Internet Protocol
JSON	JavaScript Object Notation
NER	Named Entity Recognition
NLP	Natural Language Processing
OEM	Original Equipment Manufacturer
ORDA	Online Research Data
RDBMS	Relational Database Management System
SQL	Structured Query Language
T&Cs	Terms and Conditions
TREC	Text Retrieval Conference(s)
TSV	Tab Separated Values
UDF	User-Defined Function

WP	Work Package

1 Introduction

This document describes the Data Management Plan (DMP) of the FashionBrain project. It is the initial plan that outlines what data will be generated or collected, how data will be managed, the standards in use, the workflow to make the data accessible for use, reuse and verification, and which plans for data sharing and preservation exist, ensuring that data are well-managed.

This deliverable outlines the initial DMP, which is in line with the H2020 guidelines for data management plan creation and identifies the initial classes of datasets of the project.

This DMP is not a fixed document and it covers the whole research data life cycle: creation, storage, use, sharing, archiving, destruction. The data collected or generated during the duration of the project by the consortium partners are identified as **newly collected data** in this document.

The remainder of the document is structured as follow: Section 2 presents an overview of the FashionBrain integrated architecture (FaBIAM), Section 3 summarise the data used and collected by the consortium, Section 3.1 describe the data processing workflow, Section 5 discusses the findability, accessibility, interoperability and reuse (FAIR) of the data. Section 6 presents the resources allocation for the management of the data. Section 7 discusses the security of the data, and Section 8 the ethics aspects of the data collection. Finally, Section 9 provides links to institutional and consortium policies.

1.1 Scope of This Deliverable

We refer to deliverables D9.1 to D9.4 for more information of the Ethics aspects of the data collection, and to Deliverable D2.3 for more information on the FashionBrain integrated architecture.

2 Data Integration in FashionBrain

The main output in terms of data integration (at the moment of publication of this deliverable) is FashionBrain Integrated Architecture (FaBIAM), a MonetDB-based architecture for storing, managing and analysing of both structured and unstructured data, which has a three layer structure:

- At the bottom, the *data ingestion layer* supports ingestion and storage of both structured (i.e. Comma Separated Values (CSV)) and unstructured (i.e. JavaScript Object Notation (JSON)) data.
- In the middle, the *processing layer* supports advanced Structured Query Language (SQL) window functions, in-database machine learning and continuous queries. These features have been added into the MonetDB kernel in the context of project task T2.3.
- At the top, the *analysis layer* has integrated tools provided by FashionBrain partners for advanced time series analysis (with partner UNIFR) and text data processing (with partners BEUTH and Zalando).

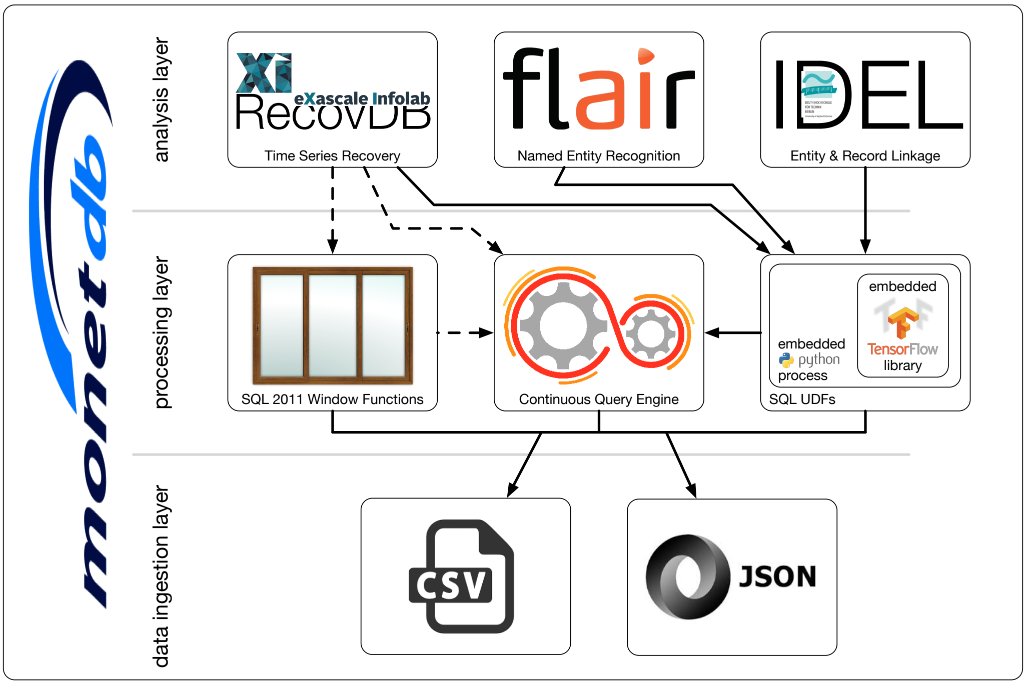


Figure 2.1: Architecture of the FashionBrain integrated architecture (FaBIAM).

In FaBIAM the following work with partners has been integrated:

- RecovDB [2]: an advanced time series missing value recovery system using MonetDB and centroid decomposition [3]. This work is reported in detail in the associated deliverable *D2.4 "Time Series Operators for MonetDB"* and has been submitted to 35th IEEE International Conference on Data Engineering (ICDE 2019). This work is in collaboration with partner UNIFR.
- In-Database Entity Linking (IDEL) [4]: an entity linking system for both text data and relational records based on neural embeddings. This work was reported in detail in deliverables *D4.1 "Report on text joins"* and *D4.2 "Demo on text joins"* and is going to appear in 6th IEEE International Conference on Big Data and Smart Computing (BigComp 2019). This work is in collaboration with partner BEUTH.
- FLAIR[1]: a Natural Language Processing (NLP) library[1] based on word and document embeddings. This work was reported in detail in deliverable *D6.2 "Entity linkage data model"* and was published in 27th International Conference on Computational Linguistics (COLING 2018). This work is in collaboration with partner Zalando.

Figure 2.1 shows an overview of FaBIAM. All components are integrated into the kernel of MonetDB. The solid arrows indicate the components that can already work together, while the dashed arrows indicating future integration. From bottom to top, they are divided into three layers:

Data Ingestion Layer This layer at the bottom of the MonetDB kernel provides various features for loading data into MonetDB. In the fashion world, there are three major groups of data: structured (e.g. product catalogues and sales information), unstructured (e.g. fashion blogs, customer reviews, social media posts and news messages) and binary data (e.g. videos and pictures). A prerequisite for the design of FaBIAM is that it must be able to store and process both structured and unstructured data, while binary data can be generally left as is. Therefore, in addition to CSV (the de facto standard data format for structured data), MonetDB also support JSON (the de facto standard data format for unstructured data) as a native data type.

Processing Layer This layer in the middle of the MonetDB kernel provides various features to facilitate query processing. In the context of the FashionBrain project, in general, and WP2 in particular, we have introduced several major extensions in this layer geared towards streaming and time series (fashion) data processing by means of both traditional SQL queries, as well as using modern machine learning technologies. This include i) major extensions to MonetDB's support for Window Function, which is detailed in the associated deliverable *D2.4 "Time Series Operators for MonetDB"*; ii) a Continuous Query Engine (CQE) for streaming and Internet of Things (IoT) data, and iii) a tight integration with various machine learning libraries, including the popular

[1] https://github.com/zalandoresearch/flair

TensorFlow library, through SQL Python User-Defined Function (UDF)s.

Analysis Layer In this layer at the top of the MonetDB kernel, we have integrated technologies of the other FashionBrain partners to enrich MonetDB's analytical features for (fashion) text data and time series data:

- *FLAIR* [1] is a python library (provided by Zalando) for named entity recognition. We refer to Deliverable D2.3 for more information on how one can use FLAIR from within MonetDB through SQL Python UDF.
- *IDEL* [4] is also a python library (provided by BEUTH), but for linking already identified entities between text data and relational records, and for records linkage of already identified entities (between relational records). The integration of IDEL in MonetDB was described in details in earlier deliverables *D4.1 "Report on text joins"* and *D4.2 "Demo on text joins"*.
- *RecovDB* [2] is a MonetDB-based Relational Database Management System (RDBMS) for the recovery of blocks of missing values in time series stored in MonetDB. The Centroid Decomposition (CD)-based recovery algorithm (provided by UNIFR) is implemented as SQL Python UDFs, but UNIFR and MDBS are working together on porting it to MonetDB native C-UDFs. This work is detailed in the associated deliverable *D2.4 "Time Series Operators for MonetDB"*.

In summary, the design of the FaBIAM architecture covers the whole stack of data loading, processing and analysis specially for fashion text and time series data. We refer to Deliverable D2.3 for more information on each layer, and to show how FaBIAM can be used to process, analyse and store a stream of reviews posted by Zalando customers.

3 Data Summary

The FashionBrain project makes use of the following kind of data: existing data from third-party sources (previously collected data), datasets generated by the consortium partners before the inception of the project (proprietary data), and data generated or collected during the project (newly collected data). A summary of each of these for each partner of the consortium is provided below, but first, an overview of the data in context of the project as a whole, is given. The FashionBrain project aims at combining data from different sources to support different fashion industry players by predicting upcoming fashion trends from social media as well as by providing personalized recommendations and advanced fashion item search to customers.

The kind of data used and collected is heterogeneous and spans from retailer data, such as sales information and item reviews, to social media information related to fashion (images and connectivity information), as will be described in detail in the rest of this section. In this project, newly collected data are self-reported data from interviews or questionnaires, unobtrusive observation/crowdsourcing (in person and/or via computer) using log files, screen capture video, as well as the text of articles, blogs, and social media posts.

Existing third-party data refers to data collected through previous research campaigns, such as CLEF, TREC, or other types of independent research, which are usually used as benchmarking/reference dataset to test new techniques. Existing proprietary data are mainly industrial data containing sensitive information such as sales data.

A quick summary of the data that has been produced by members of the consortium is provided in Table 3.1. The data collected during the duration of the project by the consortium partners are identified as **newly collected data** in this document.

The data is used by the consortium to gain knowledge on trend prediction, data integration and advanced search in the fashion domain. The remainder of this section summarises specific details regarding data composition and usage.

Table 3.1: FashionBrain Data Summary of data released by or generated by the consortium.

Dataset description	Partner	Available from	Type	Dissemination level	related WPs	License	Newly collected
Shop Inventory	Fashwell	M3	Unstructured publicly available metadata including images and text description	Public	WP2, WP4	CC-BY	no
Social Media Blogs	Fashwell	M3	Instagram profiles (posts, metadata) and 30 wordpress blogs (posts, text, images, metadata)	Public	WP4, WP5	CC-BY	no
User reviews in 11 languages	Zalando	M3	3.7mn approved and anonymised used reviews linked to 680 000 Zalando articles	Private	WP5, WP6	na	no
Product catalog data	Zalando	M3	Images of 2.3mn articles with on average 5 detail images and 11 OEM attributes per article	Private	WP2, WP6	na	no
Multilingual translations of OEM information in 11 languages	Zalando	M3	Translations of 10 700 OEM attributes in up to 11 languages	Private	WP6	na	no
Zalando Sales Data	Zalando	M3	Two years of aggregated sales data from Zalando split over seasons and commodity groups	Private	WP5	na	no
FEIDEGGER and FashionNER	Zalando, USFD	M12, M24 (enriched version)	A set of description of fashion images	Private (FEIDEGGER available on request)	WP6	na	yes
Crowdsourcing data	UNIFR, USFD, Fashwell	M3-30	An heterogeneous set containing crowdsourcing labeling and user activities (see Sections 3.3.2, 3.3.4, and 3.3.5)	Private	WP3,WP5,WP6	na	yes
Focused sampling	UNIFR	M36	graph structure of Twitter user interested in fashion, used mainly to train models	Private	WP3,WP5,WP6	na	yes
Social Media Images	Fashwell	M12	metadata of fashion influencers	Private	WP3,WP5,WP6	na	yes

3.1 Existing Proprietary Data

User reviews in 11 languages. (Industrial data)

This confidential dataset contains reviews of Zalando items in 11 different languages. The content has been anonymised (with any record containing personal data, e.g. telephone numbers, being removed). This dataset contains 3.7 millions reviews (linked to 680 000 Zalando articles) and was released in month 3 (of 36) of the project.

Relation to the objectives of the project: It will be used to improve trend predictions and to improve search (WP5, WP6).

Types and formats: The dataset is available in CSV format.

Origin: Zalando

Data utility: It will be used by UNIFR, USFD, and Zalando.

Product catalog data (Industrial data)

This confidential dataset contains images of 2.3 million articles with, on average, 5 detailed images and 11 OEM attributes per article, which was released in month 3.

Relation to the objectives of the project: It will be used in WP2 and WP6 to benchmark the data integration tools.

Types and formats: The dataset is available in CSV format, with links for the jpg images.

Origin: Zalando

Data utility: It will be used by UNIFR.

Multilingual translations of OEM information in 11 languages (Industrial data)

This confidential dataset contains translations of 10.700 OEM attributes in up to 11 languages, and was released in month 3.

Relation to the objectives of the project: It will be used in WP6 to improve multilanguage search.

Types and formats: The dataset is available in CSV format.

Origin: Zalando

Data utility: It will be used by Zalando.

Zalando Sales Data (Industrial data)

This confidential dataset contains two years of aggregated sales data from Zalando split over seasons and commodity groups, and was released in month 3.

Relation to the objectives of the project: It will be used in WP5 to improve trend predictions.

Types and formats: The dataset is available in CSV format.

Origin: Zalando

Data utility: It will be used by UNIFR and Zalando.

Shop Inventory of around 100 online shops. (Industrial data)

This dataset contains unstructured publicly available metadata including images and text description from 100 online shops. A processed version of the product categories was made available in month 3 with a CC-BY license.

Relation to the objectives of the project: It will help UNIFR with the development of the FashionBrain Taxonomy.

Types and formats: the dataset is available in JSON format.

Origin: Fashwell

Data utility: it will be used by UNIFR and Zalando, and by external researchers that work on fashion taxonomies.

Social Media Blogs

This datasets contains a list of 100 fashion influencers, along with their associated blogs. An unlinked version (without the associated blog) was publicly released in month 3 with a CC-BY license.

Relation to the objectives of the project: It will help UNIFR to study fashion trends in social media.

Types and formats: The dataset is available in tab separated values (TSV) format.

Origin: Fashwell

Data utility: It will be used by UNIFR and USFD, and by external researchers that work on fashion trends.

3.2 Existing Third-party Data

CoNLL03 NER corpus

This dataset is a small set of web pages annotated with NER data, used in the 2009 CCG NER CoNLL paper "Design Challenges and Misconceptions in Named Entity Recognition" of Lev Ratinov and Dan Roth.

Relation to the objectives of the project: It will be used in WP6 to benchmark NLP solutions for search.

Origin: https://cogcomp.org/page/resource_view/28.

Data utility: It will be used by Zalando.

CoNLL-2000 Shared Task Dataset

This dataset is a corpus of sentences with chunking: syntactically correlated parts of words.

Relation to the objectives of the project: It will be used in WP6 to benchmark NLP solutions for search.

Origin: https://www.clips.uantwerpen.be/conll2000/chunking/.

Data utility: It will be used by Zalando.

Universal Dependency Treebanks

A dataset of cross-linguistically consistent grammatical annotations.

Relation to the objectives of the project: It will be used in WP6 for multi-lingual functionalities.

Origin: https://universaldependencies.org/.

Data utility: It will be used by Zalando.

DeepFashion consumer2shop

Relation to the objectives of the project: It is used to benchmark image-attribute retrieval models.

Origin: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion/Consumer2ShopRetrieval.html

Data utility: It will be used by Fashwell.

TPC-H

Relation to the objectives of the project: It is used to benchmark MonetDB functionalities of query retrieval and data modification.

Origin: http://www.tpc.org/tpch/.

Data utility: It will be used by MDBS.

Air Traffic data (U.S. domestic commercial flight statistics)

Relation to the objectives of the project: It is used to benchmark MonetDB functionalities of data ingestion.

Origin: http://www.transtats.bts.gov/.

Data utility: It will be used by MDBS.

Reuters corpus RCV-1

Relation to the objectives of the project: It is used to benchmark MonetDB functionalities for entity linking (IDEL).

Origin: https://trec.nist.gov/data/reuters/reuters.html.

Data utility: It will be used by MDBS.

Amazon product data

Relation to the objectives of the project: It is used to benchmark fashion trend prediction (D5.3).

Origin: http://jmcauley.ucsd.edu/data/amazon/.

Data utility: It will be used by UNIFR.

Crowdsourcing Annotation Logs

Relation to the objectives of the project: It is used to improve crowdsourcing tools and techniques (WP3).

Origin: https://github.com/kbenoit/CSTA-APSR.

Data utility: It will be used by USFD.

WebNLG Dataset

Relation to the objectives of the project: It is used to benchmark entity linking techniques (D4.1 and D4.2).

Origin: Licensed by the Wikimedia Foundation http://webnlg.loria.fr/pages/challenge.html.

Data utility: It will be used by Beuth.

TACRED Dataset

Relation to the objectives of the project: from Wikimedia Foundation: it is used to benchmark stacked deep learning techniques (D4.3 and D4.4).

Origin: Released by the linguistic data consortium https://catalog.ldc.upenn.edu/LDC2018T24.

Data utility: It will be used by Beuth.

3.3 Newly Generated Data

3.3.1 FEIDEGGER and FashionNER

These confidential datasets are new multi-modal corpora developed by Zalando that focus specifically on the domain of fashion items and their visual descriptions in German. They contain a set of descriptors for 8732 fashion images obtained via crowdsourcing (in collaboration with USFD), which were released in month 12, with an enriched version released in month 24.

Relation to the objectives of the project: It will help to improve NLP and data integration solutions.

Types and formats: It is in JSON format, with links for the jpg images which are available separately.

Origin: Zalando

Expected size: 80GB.

Data utility: It will mainly used by Zalando with FEIDDEGER also being available, upon request, to external collaborators (outside the consortium).

Ethics This data collection process is following the ethics application "Crowdsourcing" (see Section 4.1 of Deliverable D9.1).

3.3.2 Annotated Images Crowdsourcing Data

Fashwell is collecting data via crowdsourcing to annotate images to improve data normalisation techniques, as described in Deliverables D3.1 and D5.1.

Relation to the objectives of the project: It will be used in WP3 and WP5.

Types and formats: The data collected are in JSON format and contain metadata on fashion images to link them to known fashion products.

Origin: Fashwell

Expected size: 10GB.

Data utility: It will mainly be used by Fashwell and USFD.

Ethics This data collection process is following the ethics application "Crowdsourcing" (see Section 4.1 of Deliverable D9.1).

3.3.3 Social Media Images

Fashwell is collecting metadata of fashion influencers to test and showcase the brand/product recognition capabilities of their technologies. This will be a confidential dataset.

Relation to the objectives of the project: It will be used in WP5.

Types and formats: The data collected are in JSON format and contain metadata of Instagram fashion influencers.

Origin: Fashwell

Expected size: <1GB.

Data utility: It will be used by Fashwell.

Ethics This data collection is following the ethics application "Social Media" (see Section 4.2 of Deliverable D9.1).

3.3.4 Annotated Reviews Crowdsourcing Data

USFD is collecting data via crowdsourcing to annotate and enrich Zalando reviews.

Relation to the objectives of the project: It will be used in WP3, specifically in D3.2.

Types and formats: The data collected are in JSON format and contain metadata on text highlights as well as associated classes.

Origin: USFD and UNIFR.

Expected size: <1GB.

Data utility: It will be used by UNIFR, Beuth and USFD.

Ethics This data collection process is following the ethics application "Crowdsourcing" (see Section 4.1 of Deliverable D9.1).

3.3.5 User Behaviour via Crowdsourcing

USFD is collecting data via crowdsourcing to study user behaviour and bias in fashion.

Relation to the objectives of the project: It will be used in WP3.

Types and formats: The data collected are in JSON format and contain logs of user interaction with web forms.

Origin: USFD and Zalando.

Expected size: <10GB.

Data utility: It will be used by Zalando and USFD.

Ethics This data collection process is following the ethics application "Crowdsourcing" (see Section 4.1 of Deliverable D9.1).

3.3.6 FashionBrain Taxonomy

UNIFR is generating a fashion taxonomy that aggregates and extends the existing open source taxonomies such as Google, ebay and Amazon taxonomies with the consortium internal taxonomy, i.e, Fashwell Shop Inventory (Section 3.1). A key advantage of the FashionBrain taxonomy, compared to existing taxonomies, is the redundancy reduction which is often caused by by gender dependency and additional classes such as occasion, sport, etc. The FashionBrain project avoid this redundancy problem by making its taxonomy gender neutral and by using additional classes, such as occasion, as a subcategory describing an item's feature. The new taxonomy is also publicly available in the from of a graphical tool that can be accessed through this link: https://fashionbrain-project.eu/fashion-taxonomy/. The tool allows users to navigate throughout all levels of the taxonomy in an interactive way.

Relation to the objectives of the project: It will be used in WP2 and WP6.

Types and formats: The data collected are in JSON format.

Origin: UNIFR.

Expected size: <1GB.

Data utility: It will be used by UNIFR and potentially by the fashion industry.

Ethics This data generation is an enrichment of existing data and does not use any personal data: for this reason it did not require an ethics application.

3.3.7 Focused Sampling

UNIFR and USFD is collecting information on the graph structured of fashion users in fashion. The goal is to train mathematical models on user behaviours and interests. No personal data will be released. The dataset used to train the models is private.

Relation to the objectives of the project: It will be used in WP3, WP5 and WP6.

Types and formats: The data collected are in JSON format.

Origin: UNIFR and USFD.

Expected size: <1GB.

Data utility: It will be used by UNIFR and USFD.

Ethics This data collection process is following the ethics application "Focused Sampling" (see Section 4.3 of Deliverable D9.1).

4 Data Processing

In this section we provide details on the workflow and methodologies followed in the data processing phase of newly collected data. Unless otherwise specified, the workflow applies to all data collection processes in the consortium. Notable exceptions are the data anonymisation and data quality processes. For this reason we include specific subsections on data quality and anonymisation, where we explore the differences between the datasets.

4.1 Workflow

The workflow of data processing of newly collected data is the following:

Anonymisation User IDs and IP addresses are removed and substituted by a random string, in a way that it is not possible to re-identify the single user. Only the metadata of the record contain personal data, so the content does not need anonymisation (this is different with, e.g., Zalando existing proprietary data, were in data cleaning phase the content has been stripped of personal information like phone numbers).

Data summarisation Analysis of word frequencies is performed and other statistics are collected.

Data aggregation Data are grouped by type (e.g. joining images from same countries).

validation and quality check Potential outliers and errors are flagged using automated techniques.

Data cleaning Ouliers and errors are removed.

Data analysis Models are created using the data as training set (e.g. neural network training set).

4.2 Data Quality

The data quality process is based on the following principles:

Data range Check for and correct values out of range.

Validation Check the entered data against a few randomly selected ground truth data points, that have been carefully created by the researcher.

Dimension sanity check Check the lengths of rows and the number of variables against the planned values.

Robust naming Do not recode variables in the data collection phase.

Reconstructible workflow After the data have been collected, a back-up file and a separate working copy should be created. Moreover, is is important to be consistent on the representation of missing data and error values.

Documentation Create documentation of all changes made to the dataset in a way that the data processing could be reproduced by an external researcher. Jupyter notebook files are used to guarantee that the coding and the intended output are coupled together.

Variable Recoding

During the analysis of the data, there could be the need to recode variables (for example grouping by region), or to form new variables from the existing ones. All such changes made to variables will be well-documented.

Missing Data

In almost all datasets, there are variables with missing data.

The missing data will be coded so that they can be clearly distinguished from the other values of a variable. Often, values such as 9, 99, 999, or 0 are used to signify missing data. We discourage this approach, and the use of "NaN" or an empty field is encouraged in this project.

Weight Variables

If there are systematic errors in the dataset, it is appropriate to weigh the observations. This is especially true for the three crowdsourcing data collection processes. With weight variables, potential bias in age, gender and region distributions resulting from the sampling can be corrected. Clear documentation on the weighting methods and calculations used will be provided to ensure that people reusing the data also have an understanding of the variables created during the research process.

4.3 Anonymisation

The anonymisation process of the various data collection processes is explained below.

Crowdsourcing datasets (3.3.2, 3.3.4, and 3.3.5), FEIDEGGER and FashionNER The raw input datasets are log files from crowdsourcing platforms, containing personal metadata: user ID and IP address, and in some case a user alias (that could contain personal data). The anonymisation process consists in deleting the IP address and the user alias, and in substituting the user ID with a random number, in such a way that re-identification is not possible. The collection phase can be long in some cases (longitudinal studies) and for this reason the members of the consortium will have three months to complete the collection phase, and must anonymise the data after that. The participants are aware of this procedure, as specified in detail in Deliverable D9.1 and in this document in Section 8, and as documented in the consent forms shown in Deliverable D9.4.

FashionBrain Taxonomy No personal data are collected so no anonymisation is needed.

Social Media Images The raw input dataset consists of Instagram metadata of public fashion influencers and contains the following personal data: user IDs, image links, text, user alias. User IDs and aliases will be deleted, and only the image links (not the actual images) are stored.

Focused Sampling The raw input dataset consists of crowdsourcing logs from Twitter users, and as such it contains personal data: crowdsourcing user IDs, IP address, user aliases, and Twitter handles. The anonymisation process consists in deleting the IP address and the user alias, and in substituting the user ID with a random number, in such a way that re-identification is not possible. After the data collection phase is completed, the twitter handles that could allow identification of users, or of the users' graph structure will be substituted with random numbers, and only the mathematical model representing aggregated user behavioural models will be retained. The participants are aware of this procedure, as specified in detail in Deliverable D9.1 and in this document in Section 8, and as testified by the consent forms in Deliverable D9.4.

5 FAIR Data

5.1 Making Data Findable

Discoverability of Data (Metadata Provision)

Metadata provided will be minimal (type of dataset, scope and format), as the datasets provided are for research use, and thus, are idiosyncratic and non-standardised in nature. In this type of research discipline, there are no metadata standards. However, the project will strive to follow the de facto standards created by previous datasets in the field (e.g. DeepFashion http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html).

Identifiability of Data

All data that has been publicly released will be assigned a permanent identifier through EUDAT (https://trng-b2share.eudat.eu/), which can be retraced to the University of Sheffield.

Naming Conventions

The naming convention used will be consistent with the existing datasets in the field. For example, the crowdsourcing data will use the standard Amazon Mechanical Turk (AMT) convention and image data will use the same convention of DeepFashion. These examples, along with other cases, are illustrated in Table 3.1.

Keywords

Keywords used will be related to the type of dataset (e.g., fashion, catalog) and the application (e.g., taxonomy).

Versioning

Versioning will follow EUDAT standards where a new version of the dataset can be created containging the same metadata but under a different file. For these new records, new persistent identifiers and file storage locations are used with MAJOR.MINOR numbering system.

5.2 Making Data Openly Accessible

The following datasets will be publicly released and available (as explained in detail in Section 3.3):

Social Media Blogs This dataset contains a list of 100 fashion influencers, along with their associated blogs. An unlinked version (without the associated blog) was publicly released in month 3 with a CC-BY license.

Shop Inventory of around 100 online shops This dataset contains unstructured publicly available metadata including images and text descriptions from 100 online shops. A processed version of the product categories was made available in month 3 with a CC-BY license.

FashionBrain Taxonomy The consortium developed a new taxonomy to serve as a backbone knowledge base for all partners involved in the FashionBrain project. The new taxonomy aggregates and extends the existing open source taxonomies such as Google, ebay and Amazon taxonomies with the consortium internal taxonomy, i.e, Fashwell taxonomy.

FEIDEGGER This dataset is a new multi-modal corpus that focuses specifically on the domain of fashion items and their visual descriptions in German. It contains a set of descriptors for 8732 fashion images obtained via crowdsourcing. As explained in Section 3.3, it is only available upon request (to external researchers) to protect the project's intellectual property rights and possible issues of competition during the research and development stages.

The first three datasets will be released through EUDAT, and a metadata-only record will be deposited in the ORDA hub, referring back to the EUDAT version of the record.

The industrial data described in Table 3.1 will be not made publicly available as they are confidential in nature (sales figures, etc.) and would adversely affect the businesses of the partners.

The crowdsourcing data and Twitter data will not be made publicly available as the information useful to the research community is not the data itself, but rather the trained models and the technologies built using this data. For this reason, FashionBrain has created a Github repository (https://github.com/FashionBrainTeam) where all the non-proprietary and non-commercial algorithms are published. Github is integrated with EUDAT, guaranteeing a wide accessibility.

No additional or proprietary methods or software tools are needed to access the available data, as they will be in standard JSON, CSV, or TSV formats.

There will be no licensing restrictions when accessing the publicly available data (see Section 3.1 for further information).

5.3 Interoperability

To achieve interoperability within the consortium, the project has developed an integrated data architecture, as described in Section 2.

As explained in section 4.1, metadata provision will be minimal for the released data as the datasets provided are for research use, and thus, are idiosyncratic and non-standardised in nature. In this type of research discipline, there are no metadata standards. However, the project will strive to follow the de facto standards created by previous datasets in the field (e.g. DeepFashion http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html.

The naming convention used will be consistent with other existing datasets in the field as shown in Table 3.1.

5.4 Data Reuse

- The datasets will be licensed with Creative Commons Attribution (CC-BY)[1] with the following requirements:

  Attribution You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

  No additional restrictions You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

- There will be no embargo.
- The data will be released in the EUDAT platform indefinitely.
- The data quality will be monitored and maintained until the end of the project (31 December 2019).

[1] https://creativecommons.org/licenses/by/4.0/

6 Allocation of Resources

6.1 Cost

The cost of making research data FAIR is eligible for reimbursement during the duration of the project under the conditions defined in the H2020 Grant Agreement[1], in particular Article 6 and Article 6.2.D.3, but also other articles relevant for the cost category chosen.

Costs are most likely be incurred when ensuring open access to articles and other data on publisher's websites and/or hosting platforms. These costs are anticipated to be in the region of 3000 € over the lifetime of the project.

The person months to maintain FAIR data practices are included in the USFD administration person months of WP8 (3 person months of the 15 allocated for that work package).

6.2 Data Management Responsibility

USFD is responsible for the data management of all data that are publicly released. Proprietary data, as described in Section 3, are managed by the respective industrial partner.

6.3 Costs and Value of Long Term Preservation

The cost for long term preservation of publicly available data is minimal, as it is managed via the EUDAT B2Share platform which is free of charge for European scientists and researchers.[2]

Preservation of non-public data cannot be costed as it is commercially sensitive and therefore accessible only while the project is active.

All of the consortium data are stored in a password-protected Google Drive Enterprise account that has an integrated backup solution. The cost of this solution is covered by USFD.[3]

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] https://eudat.eu/services/userdoc/b2share#What_about_Costs_and_Trust

[3] USFD are satisfied that the security controls put in place by Google are sufficient to protect University data. A summary of their assessment can be found at: https://www.sheffield.ac.uk/polopoly_fs/1.254519!/file/FactSheet-DataSecurityandPrivacywithGoogle.pdf

7 Data Security

We now summarise the data security policies in place for the FashionBrain project.

Data recovery Data recovery will be operated by the University of Sheffield Corporate Information and Computing Services, that is using a Google Drive enterprise account with backup solutions in place.

Data transfer Data will be transferred through secure file system protocol (SFTP).

Long-term storage Long term storage will be on discretion of the University of Sheffield.

Processing The data recorded will be encrypted and securely stored on password protected computers at the Data Controller and a copy may also be stored on other members' laptops for analysis purposes. The encrypted and password protected data will be backed up on an external drive kept in a locked drawer in the data controller's safe storage unit.

8 Ethical Aspects

We refer to Deliverable D9.3 for an in detail analysis of the ethical aspects of the data collections performed during the duration of the project, and to Deliverable D9.4 for the consent forms signed by the participants. For a description of the anonymisation process we refer to Section 4.3 in this document. We summarise here the most relevant points.

8.1 Consent

For the majority of newly generated data, informed consent will be obtained from the participants, except in the data collection of "Social Media Images". The participant will have the opportunity to leave the task at any time and can request to have their personal data removed for up to 3 months after its collection[1]. The participant will have to sign the form (or click the "I agree" button in the online version) in order to perform the task.

For the data collection "Social Media Images", only implicit consent will be obtained[2], and we refer to Section 8.4 for a more detailed analysis of consent for this action.

8.2 Erasure

In the sole case of "Social Media Images" data collection (Section 3.3), in which public data are obtained without consent, Art. 17 of GDPR (right to erasure) should be guaranteed. For this reason, the EU reviewers and the FashionBrain Ethics Committee recommended the development of a Consent Manager which is available on the project website as a tool that allows the public to request an opt-out from FashionBrain data collection. We refer to Section 8.4 for more information on the Consent Manager.

For all the other data collection processes, the consent form will contain (in addition to the project information, data use and privacy issues) the email contact address to withdraw consent and request erasure of the data so long as it is within 3 months after the collection phase. After this point, the data will have been anonymised and it will no longer be possible to identify the data in order to delete it.

[1] This finite window of time is necessary as personal identifiers will be destroyed after that period.

[2] users agreed to the T&Cs of the social media platform they are using about access of their data via the platform Application Programming Interface (API).

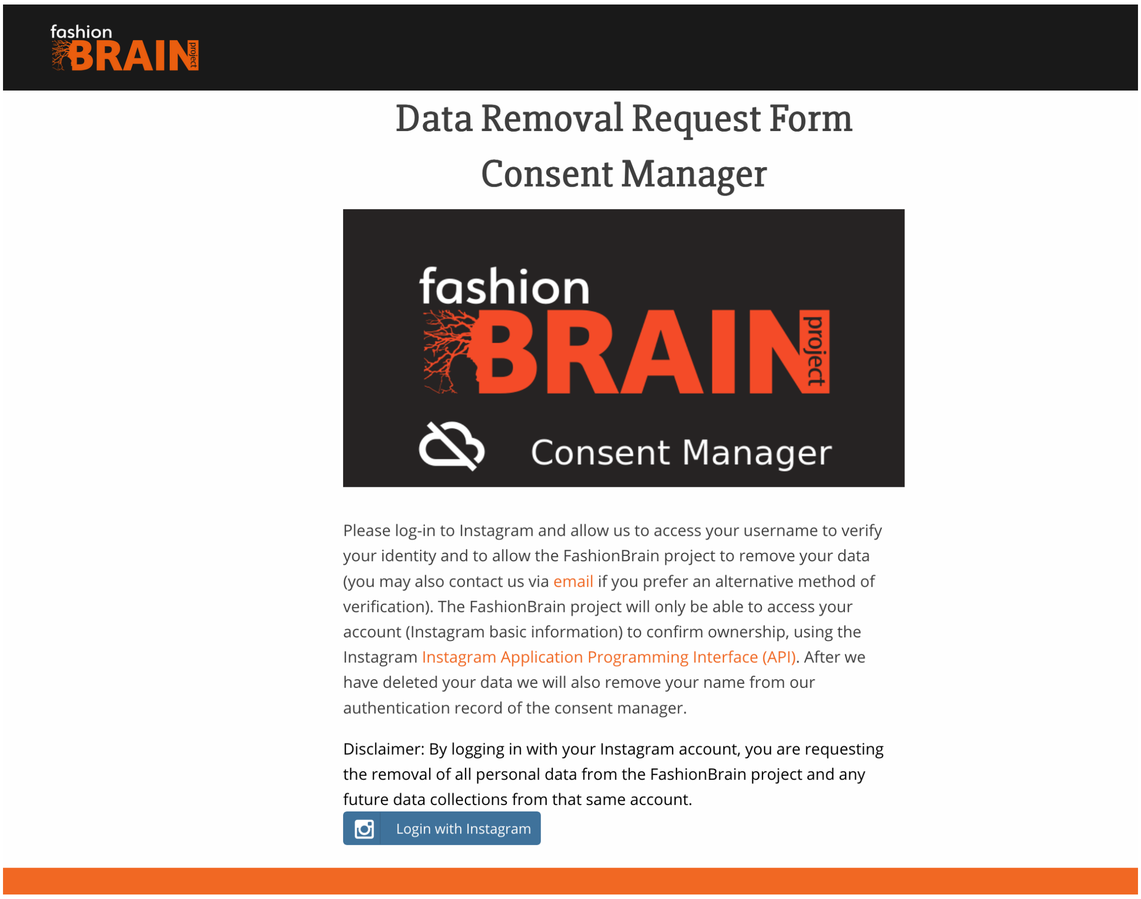
8.3 Data Sharing

The participants give explicit permission for the research team to re-use the data for future research.

8.4 Consent Manager

The Consent Manager has been published on the project website: https://fashionbrain-project.eu/consent_manager (Figure 8.1). It uses the Instagram Application Programming Interface (API) to verify the identity of the data owner that is making the opt-out request from FashionBrain data collection and usage, without the risk of disclosing any other personal data. Technically, this works by redirecting the authentication to Instagram through the API and subsequently receiving a token back (`access_token`) from Instagram that allows minimal access to the Instagram profile (`ID`, Instagram name). A request sent through this page will activate the data removal and blacklisting procedure. From the date of receipt of the opt-out request, FashionBrain will destroy all past records belonging to the requester, and the account ID will be added to a static blacklist table to prevent future data collection from that account.

As previously explained, the Consent Manager is in place for cases in which public data is obtained without consent from the FashionBrain Project. Instead, implicit consent has been given on another platform when the users agree to its T&Cs. For that reason, the owners of the data will not be personally notified of the existence of the Consent Manager. However, should they wish to have their current data or any future data removed, they will be able to do so via the website's Consent Manager. The consortium will advertise this page to reach the widest audience possible. Along with the Consent Manager on the website, a privacy page has been added to explain what kind of public data is being collected and what is being done with this data, to ensure adherence to the principles of fairness and transparency of GDPR (https://fashionbrain-project.eu/data-ethics-and-privacy).



Data Removal Request Form

Consent Manager

Please log-in to Instagram and allow us to access your username to verify your identity and to allow the FashionBrain project to remove your data (you may also contact us via email if you prefer an alternative method of verification). The FashionBrain project will only be able to access your account (Instagram basic information) to confirm ownership, using the Instagram Instagram Application Programming Interface (API). After we have deleted your data we will also remove your name from our authentication record of the consent manager.

Disclaimer: By logging in with your Instagram account, you are requesting the removal of all personal data from the FashionBrain project and any future data collections from that same account.

Figure 8.1: FashionBrain Consent Manager.

9 Other

Below is a summary of associated institutional, national and international policies that guide the project. We refer to Deliverables D9.1 and D9.3 for more information on the specific FashionBrain policies for data protection and ethics.

9.1 University of Sheffield Institutional Policies

Data Management Policy & Procedures

University of Sheffield Research Data Management Policy http://www.shef.ac.uk/polopoly_fs/1.553350!/file/GRIPPolicyextractRDM.pdf.

Data Security Policies

University of Sheffield Data protection policy https://www.sheffield.ac.uk/govern/data-protection.

Data Sharing Policy

http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm.

Institutional Information Policy

University of Sheffield Good Research and Innovation Practice (GRIP) Policy http://www.sheffield.ac.uk/polopoly_fs/1.356709!/file/GRIPPolicySenateapproved.pdf.

Institutional Ethics Policy

The University of Sheffield Ethics Policy Governing Research Involving Human Participants, Personal Data and Human Tissue https://www.sheffield.ac.uk/rs/ethicsandintegrity/ethicspolicy/index.

9.2 Partner Policies

We refer to Deliverable D9.1 for more information on partners' national legislations and ethics policies.

MonetDB Solutions B.V. and Zalando SE are European entities and accordingly, are subject to the European data protection laws and regulations. Fashwell AG is subject to the Swiss data protection laws and regulations which afford a level of data protection equal to the European data protection laws and regulations. Accordingly, all participants will conform to and will process any personal data collected in the course of the project in compliance with (i) the applicable European data protection framework and (ii) in addition with the local law requirements, in particular (as regards Zalando SE) with the German data protection law which amongst all European member states represents one of the most strict data protection frameworks in Europe.

Bibliography

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[2] Ines Arous, Mourad Khayati, Philippe Cudré-Mauroux, Ying Zhang, Martin Kersten, and Svetlin Stalinlov. RecoveDB: accurate and efficient missing blocks recovery for large time series. In *Proceedings of the 35th IEEE International Conference on Data Engineering (ICDE 2019)*, April 2019. Submitted.

[3] Mourad Khayati, Michael H. Böhlen, and Johann Gamper. Memory-efficient centroid decomposition for long time series. In *IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014*, pages 100–111, April 2014. doi: 10.1109/ICDE.2014.6816643. URL https://doi.org/10.1109/ICDE.2014.6816643.

[4] Torsten Kilias, Alexander Löser, Felix Gers, Ying Zhang, Richard Koopmanschap, and Martin Kersten. IDEL: In-Database Neural Entity Linking. In *Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, Feburary 2019. To appear.