V4Design

Visual and textual content re-purposing FOR(4) architecture, Design and virtual reality games

H2020-779962

D2.1 Initial visual and textual dataset creation and legal and ethical requirements

Dissemination level:	Public
Contractual date of delivery:	Month 10, 31/10/2018
Actual date of delivery:	Month 10, 31/10/2018
Work Package:	WP2: Multimedia data crawling and collection for re-use and repurpose
Task:	T2.2: Movie and documentary data collection T2.3: Artwork data collection and retrieval T2.5: Ethical and legal consideration in relation to the use cases
Type:	Report
Approval Status:	Final version
Version:	1.0
Number of pages:	80
Filename:	D2.1_V4Design_InitialDatasetCreationAndLegalEthicalRequirements_20181031_v1.0.pdf

Abstract

This deliverable describes the creation of the initial multimedia and multilingual datasets for the use cases. It describes the high-level methodology that WP2 adheres to throughout the duration of the project, describes how this deliverable fits into that methodology, and lists how data was collected, provided, and evaluated throughout the last ten months. Furthermore the deliverable

contains a comprehensive description of the legal and ethical requirements.

co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	17/08/2018	ToC creation and initial content preparation	Jolan Wuyts
0.2	17/09/2018	Incorporating contributions from partners	Jolan Wuyts
0.3	28/09/2018	Inviting reviews from EF	Antoine Isaac, Victor-Jan Vos, Jolan Wuyts
0.4	10/10/2018	1st integrated draft prepared and circulated to WP2-related partners for review	Jolan Wuyts
0.5	16/10/2018	Addition of a chapter on legal and ethical requirements by DW	Nicolaus Heise (DW)
0.6	22/10/2018	Pre-final draft sent for internal review	Jolan Wuyts
0.7	28/10/2018	Internal review	Maarten Vergauwen (KUL)
1.0	30/10/2018	Preparation of the final draft	Jolan Wuyts

Author list

Organization	Name	Contact Information
EF	Jolan Wuyts	jolan.wuyts@europeana.eu
EF	Antoine Isaac	antoine.isaac@europeana.eu
AF	Kriszta Doczy	kdoczy@artfilms-digital.com
SLRS	Jesper Wachtmeister	jesper@solarisfilms.se
DW	Eva Lopez	eva.lopez@dw.com
DW	Stephan Gensch	stephan.gensch@dw.com
DW	Nicolaus Heise	nicolaus.heise@dw.com
CERTH	Spyridon Simeonidis	spyridons@iti.gr
CERTH	Konstantinos Avgerinakis	koafgeri@iti.gr

Executive Summary

This deliverable has the goal to report on the activities done towards the creation of the initial dataset for the V4Design project. The initial dataset is delivered together with this deliverable document, and is connected to **"*D3.1. Empirical study of textual data related to visual content*".**

This deliverable lists the methodology and workflow of the creation of the initial dataset by all contributing consortium partners: AF, DW, EF, SLRS as well as CERTH which is responsible for the data scraping. It also outlines the connection between consortium partners and external sources (like Flickr) and external partners (like SketchFab) in the generation of the initial dataset.

From the high-level framework of CRISP-DM it breaks down the steps of each iteration that ultimately led to the initial dataset. The requirements for the creation of the dataset are listed as Content User Requirements (CURs) and Content Technical Requirements (CTRs), together with the requirements stemming from the Pilot Use Cases (PUCs). The evaluation phase between each iteration is also recounted.

It describes the content of the datasets, both as high-level descriptions and as detailed lists. An extensive overview of the legal and ethical requirements is also given, covering every possible scenario of content delivery in V4Design. This deliverable concludes with describing the future work that will be done towards the delivery of the final dataset.

Abbreviations and Acronyms

VR	Virtual Reality
PUC	Pilot Use Case
SO	Scientific Objective
TO	Technical Objective
GLAMs	Galleries, Libraries, Archives, Museums
CHO	Cultural Heritage Object
JSON	Javascript Object Notation
RDF	Resource Description Framework
EDM	Europeana Data Model
FTP	File Transfer Protocol
CRISP-DM	CRoss-Industry Standard Process for data mining
CC BY	Creative Commons attribution
CC BY-SA	creative commons attribution - sharealike
PDM	Public Domain Mark
PUC	Pilot Use Case
CC0	No Rights Reserved
LOD	Linked Open Data
CUR	Content User Requirement
HLUR	High-level User Requirement
CTR	Content Technical Requirement

Definitions[1]

Content: A physical or Digital Object that is part of Europe's cultural and/or scientific heritage, typically held by a Data Provider. Note: You can use the terms Content and Digital Objects interchangeably.

Collection: We use the noun data collection here to describe a homogeneous set of digital cultural heritage objects that are tied together with a common theme or purpose. A collection is the equivalent of a dataset.

Digital Object: A digital representation of an object that is part of Europe's cultural and/or scientific heritage. The Digital Object can also be the original object when born digital.

Iteration: One iteration is one version of the initial dataset created for this deliverable.

Record: A record is a single manifestation of a digital cultural heritage object. This can, for instance, be a single image and a JSON file with the metadata of that image.

Freely Re-usable Content: Digital Objects that are available for re-use with minimal or no conditions, specifically those objects labelled Public Domain, CC0, CC-BY and CC-BY-SA.

Metadata: The textual information and hyperlinks that serve to identify, discover, interpret and/or manage Content. Note: This is a general term used to describe any element of Metadata.

Metadata Field: A single element of a Metadata Record describing the Digital Object.

Guideline: e.g.: 'edm:Provider' is the Metadata Field that is used to describe the Provider of the digital object.

Re-Use: The ability to make use of a Digital Object or Metadata that is available online, through the acts of sharing, duplicating, modifying or publishing.

Rights Statement: A statement that describes the conditions for Access and Re-use of Digital Objects and their Previews. Rights statement are communicated via the 'edm:rights' Metadata Field as defined by the Europeana Data Model.

[1] Most of these definitions were taken from the Europeana Glossary of Terms: https://pro.europeana.eu/resources/standardization-tools/glossary, last visited on 19/10/218

Table of Contents

1 INTRODUCTION

This deliverable outlines the workflow and outcomes of WP2 in the first ten months of the V4Design project. It gives an overview of the high-level workflow for WP2 spanning the complete length of the project, using the Cross-industry standard process for data mining (CRISP-DM) as its guiding process.

The current introductory section outlines the way of working between the partners from the start of the project until the delivery date of this deliverable.

Section 2: Methodology first situates this deliverable in the CRISP-DM process, and then breaks the current work phases down into the several iterations that went into the creation of the initial dataset. A rundown of all consortium partners that provided data to the dataset is given, followed by an overview of the actual work process that went into providing data to V4Design. A short definition of every PUC is given as a sidebar, since this is important in understanding the rest of the deliverable.

Section 3: Requirements and evaluation first references D7.2, where the High-Level User requirements (HLURs) and derived User Requirements (URs) are listed. The requirements that are relevant to the provision of content to the V4Design consortium are expanded upon and recorded in *Appendix A*. The chapter follows to chronicle the adjustments made to the content in the next iterations and the process of delivering that content. The evaluation form that was used to gather feedback can be found in *Appendix E*.

Section 4 and 5 focus on cataloguing the content of data collected in each iteration. A short description of goals and requirements per iteration is given, followed by an overview of the data. Section 4 does this for data provided by content provider partners, while section 5 is devoted to data from external sources and external partners. A full list of datasets constituting every iteration can be found in *Appendix B*.

An overview of the legal and ethical requirements in relation to the use cases is given in *Section 6: Legal and Ethical requirements*. It distinguished between Privacy and data protection, which is covered first in the section, and Copyright Law. It finishes with defining other intellectual property that wasn't covered in the previous subsections.

This deliverable concludes with outlining how the initial dataset interfaces with the other scientific and technical objectives of the V4Design project, how the tasks that led up to the creation of the initial dataset will be integrated and automated in the V4Design system architecture, what future work needs to be done towards deliverable ***"D2.3 Final visual and textual dataset creation and legal and ethical framework"*,** lessons learned, and challenges ahead.

2 METHODOLOGY

This section first describes the high-level three-year process that we're going through to provide data to the V4Design consortium, and, ultimately, the V4Design platform. The cross-industry standard process for data mining (CRISP-DM) is the template off of which this high-level process is based. It introduces the content provider partners, as well as the PUCs. This section gives an extensive overview of the steps in this process that are the focus of this deliverable. It outlines how and why different iterations of an initial dataset were created to reach sufficient business understanding and data understanding of the V4Design project.

2.1 High-level methodology

The work done in WP2 can be seen as a three-year data collection task for a wider data mining project (where the actual data mining is done in other V4Design WPs): identifying, collecting and presenting large amounts of digital cultural heritage data for re-use and repurposing by architecture, design and VR gaming creative industries. Our work goes further than mere data querying and collecting: the aim is to present the most relevant data in the format most usable by the target audiences, so the barrier for discovery and reuse is as low as possible. What is seen as 'relevant and usable' is defined by the User Requirements, which are broken down into Content User Requirements and Content Technical requirements for Work Package 2 (*Section 3: Requirements and evaluation*). The goal is also to transform cultural heritage data to make it easier to reuse. WP4 aims to extract 3D objects from cultural heritage visual data, tag those objects with meaningful keywords to increase discoverability, and segment existing 3D objects into their constituting parts so they can be remixed by users into completely new 3D models. WP3 analyses textual data associated with these 3D objects to increase its quality, and looks at larger descriptions of design objects with the goal to summarise them and enrich existing 3D objects with accurate short textual descriptions. Both WP3 and WP4 use the data that WP2 provides to train machine learning algorithms to output new objects and metadata that will be used in the V4Design platform. Lastly, WP5 integrates the different data descriptions and representations of 3D objects, interconnecting them with source data, by the aid of a semantic representation model. In that sense the work done by WP2 are the first steps in a data mining process. The complexity of this work warrants a framework in which tasks can be situated and partners feel like they know what steps they have to take.

Initial desk research preceding the start of WP2 was conducted with the goal of providing a framework for the efforts in WP2. Following a comparative analysis of framework (Azevedos & Santos, 2008), (Kurgan & Musilek, 2006), the CRISP-DM process model was chosen as the high-level process that would dictate the workflow of WP2.

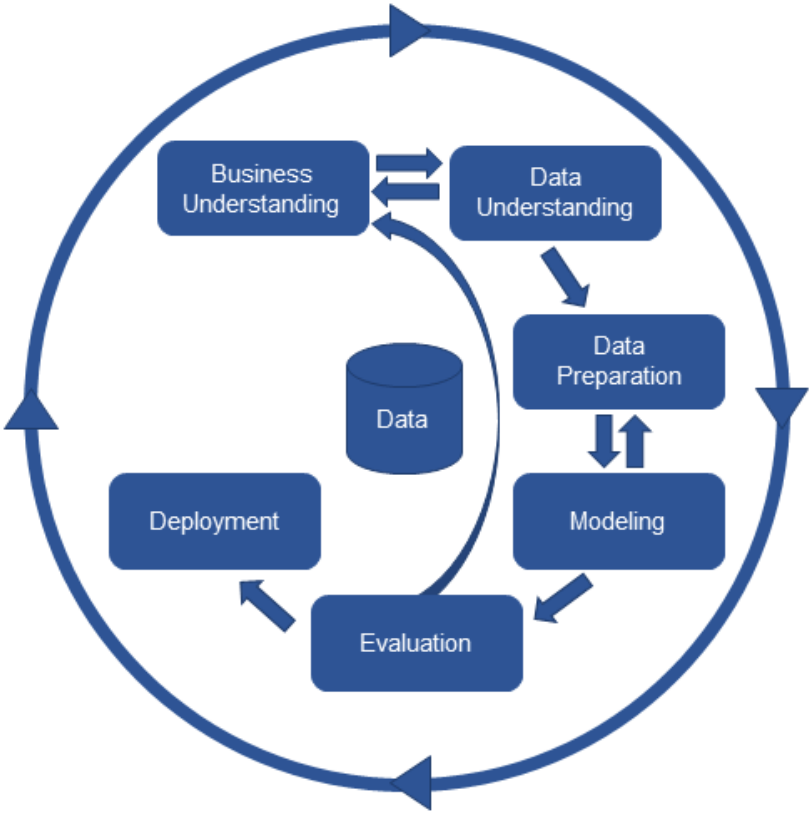

Figure 1: The CRISP-DM phases[2]

CRISP-DM provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks, and their outputs (Rüdiger, 2000). The work done in the first year of the project, described in D2.1, is situated in the first two phases of the process model: The feedback loop between business understanding and data understanding.

"Business Understanding: This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives"[3]

"Data Understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. "[4]

With this high-level overview in mind, the rest of this deliverable describes the tasks and workflow of the Business Understanding and Data Understanding Phases in more detail. The following section focuses on the low-level workflow for the Data understanding and Business understanding phases.

[2] KDNuggets, James Taylor, last accessed 28/09/2018 https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html
[3] Rüdiger 2000, 33.
[4] Ibid.

2.2 Introduction to the iterative agile workflow process for the creation of the V4Design dataset

In the V4Design DoA, the development cycle of the V4Design project is described as having four iterating cycles, with each cycle consisting of the following steps:

1. define user requirements and use cases
2. design and develop the technical architecture and platform's components
3. implement prototypes of that architecture and the platform components of step B
4. evaluate each prototype in specific use case scenarios by the architect and video game community, respectively.

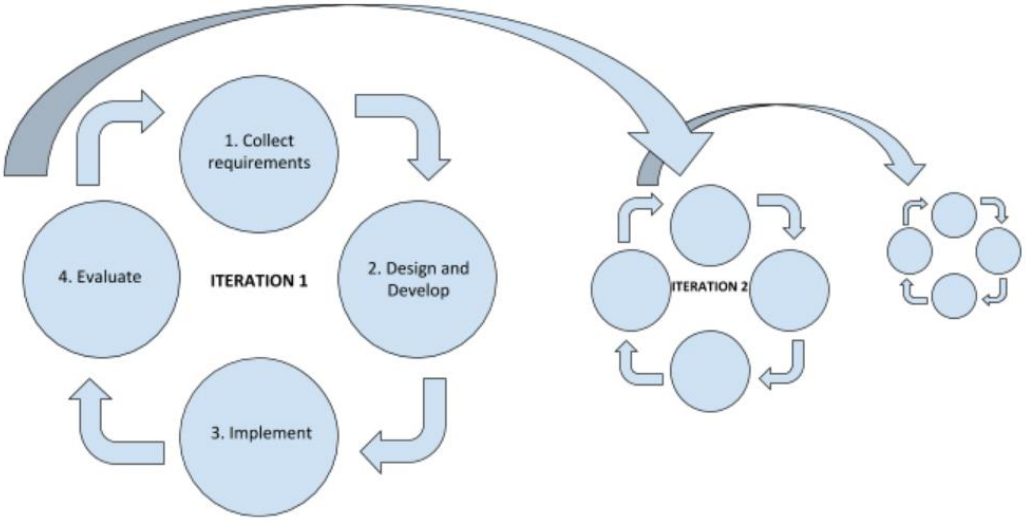

Figure 2: Diagram detailing the iterative process of dataset creation

For the creation of D2.1, we went through two iterations of this development cycle. Below is a Gantt Chart of the high-level workflow to realise D2.1:

WP2	Months	1	2	3	4	5	6	7	8	9	10
1	collect requirements for iteration 1										
2	design and develop iteration 1										
3	implement iteration 1 in other task prototypes										
4	evaluate iteration 1										
5	expand query & requirements for iteration 2										
6	design and develop iteration 2										
7	implement iteration 2 in other task prototypes										
8	evaluate iteration 2 - Delivery of the initial dataset										
9	write and deliver D2.1.										

Table 1: Gantt Chart of the process tasks executed in WP2 up until M10. Milestone months are highlighted in orange: The deadline for delivery of initial dataset iteration 1 in M6, and the deadline for delivery of the final initial dataset in M10.

Firstly, creating multiple iterations of the final dataset ensures proper evaluation of the data by consortium partners and adjustments to the data collection process where needed. We completed the first iteration in M6, in time for the milestone meeting in Bonn. At this meeting we asked every consortium member to evaluate the first iteration. That feedback was taken into account in the creation of the second iteration, which resulted in the final initial dataset included in D2.1.

Secondly, creating multiple iterations brings the deadline for delivery of a first dataset forward, which ensures that consortium partners get data for inclusion in their prototypes earlier rather than later. If component A and B of the development cycle (collecting data requirements and extracting data, respectively) take too long, this has the risk of delaying the start of tasks in other WPs that depend on WP2 reaching component C of the development cycle.

The rest of the phases correspond with the rest of the workflow envisioned in WP2, to be carried out from M10 to M32. Data Preparation will be carried out first and reported on in D2.3. The Modelling and Evaluation phases will be carried out in parallel with the Data Preparation phase, using the initial dataset. Lastly, the Modelling and Evaluation phases will be done again for the final dataset, and will feed into the Deployment phase.

2.3 Data providers

Data collection was done by the four 'content provider' consortium partners: **DW, AF, SLRS and EF**. Additionally, **CERTH** also provided a substantial amount of content, even though they are not technically a content provider partner. Their content was gathered for ***"T2.1: Web crawling and retrieval of textual and multimedia data".*** The data they delivered was almost exclusively gathered through crawling and scraping the Web, delivering content from external providers like Wikipedia, Flickr, Twitter, etc. Data gathered from T2.1, as well as data gathered from other external sources like Scantheworld and Sketchfab, is described extensively in *Section 5*: *Data collection from external providers*.

Every content provider focuses on their own types of data, making their contributions to V4Design diverse. Below is a rundown of the contents of the databases every content provider partner they agreed to share with the V4Design consortium, as outlined in the DoA.

Deutsche Welle's databases contain mostly images, text and video fragments from recent events. They provide an API[5] that is publicly available, giving access to content from the 'DW - (Breaking) World News' platform (articles, video, audio and image galleries). Additionally, content metadata and technical metadata about these data objects are also described, such as date of publication, image resolution, video length, etc. API results (data objects like image files as well as metadata) are in JSON format.

Secondly, Deutsche Welle produces a telenovela linked to a course on learning German, called Nico's Weg. One of the identified PUCs aims at using content from this telenovela to create an interactive VR learning tool. DW provides the video footage of this telenovela on their YouTube channel, as well as the text scripts and Web-scraped text from the interactive lessons for the textual semantic extraction and potential reuse in the VR learning tool.

[5] https://www.dw.com/api/config/init

ArtFilms focuses on full-length documentaries and arthouse films from a plethora of different producers and sources. The visual material provided addresses the need for collecting culturally diverse data from Europe, Asia and South Pacific.

Solaris Filmproduktion AB provides their entire catalog of movie material for re-use and research in the V4Design consortium.

Europeana Foundation provides access to over 50 million items of digitised cultural heritage from GLAMs from all over Europe. It acts as the coordinating partner for the delivery of data to the V4Design project, and contributes to the delivered data as well. Europeana provides two big types of information to the V4Design project: content, and metadata. Content as a term is used to describe the CHOs themselves, the digital representations of cultural heritage. This data may be a JPG file of a photograph of a historical painting, a WAV file of a sound recording of a piece of music, a PDF file containing the text of an old book. Metadata indicates information that describes the content, gives more information about the representation of the cultural heritage object or about the CHO itself. Europeana's metadata is always textual, and is represented in RDF or JSON format. It takes form of text fields, e.g. the field 'creator', and a textual value for that field, e.g. the value 'Leonardo da Vinci'. The combinations of fields and values are modelled in EDM. Content as well as metadata is available through Europeana's RESTful API endpoint[6]. For extraction of objects for the V4Design project both the Search and Records APIs will be used.

Apart from what is already available on Europeana Collections and the Europeana APIs, EF provides the entirety of the freely reusable Newspapers Collection handed to Europeana from The European Library project, which gives access to millions of digitised and OCRised historical newspaper pages from all over Europe, in different languages. During the first year of the V4Design project the Newspapers Collection is not available through the Europeana APIs yet, but the V4Design consortium gets exclusive access to the Newspaper Collection on the Europeana servers.

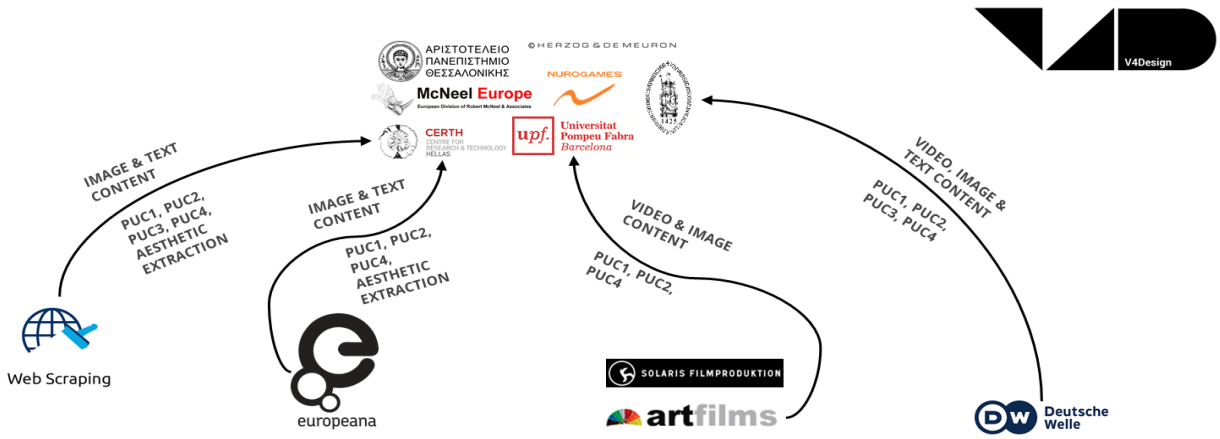

Figure 3: A schematic of the data streams from content provider partners to the other consortium partners

[6] https://www.europeana.eu/api/v2/

2.4 High-level content provision process in V4Design

At the outset of the project, V4Design created four PUCs that the V4Design platform aims to support. These PUCs are used as a guideline for what the Minimum Viable Product (MVP) should possess, both with regards to functional features as to content. These PUCs were used in M1-M4 of the project to outline what is considered 'relevant' data. Lists of keywords were set up for each PUC to define what data would be relevant. Content provider partners queried their databases using these keyword lists and extracted objects that would fit the PUCs. These keyword lists can be found in *Appendix C*. Below is a short description of each PUC, paraphrased from the DOA (pages 12-14). The next subsection describes the general step-by-step process for content delivery by each partner.

2.4.1 PUC overview

PUC 1: Architectural design, related to existing or historical buildings and sites and their environments

Scenario 1, Outdoor design for an historical landscape in Delphi, Greece: A team of architects is assigned to design a pavilion, material for the scenery and other spatial propositions for a theatrical play dated in the Late Renaissance period. These cultural heritage objects will be used to build 3D models of historical environments and will inform the design of the pavilion.

Scenario 2, Concept design for a new building in central Berlin, Germany: an architecture firm is assigned to re-design and re-furbish a landmark building, which is integrated into the grown urban fabric and possesses strong interrelations with the surrounding buildings, as well as their urban, cultural and historical attributes. This will be implemented by the architects using V4Design to search for enhanced 3D models, images and maps of the immediate vicinity and relevant reference models to study various design options.

Figure 4: Delphi - initial testing

PUC2: Architectural design, related to artworks, historic or stylistic elements

Exhibition elements on East Asian, Japanese Visual Culture: Designers are commissioned to design the interior and specific elements for an exhibition space, (inside a museum project designed by HdM), so as to host a temporary exhibition on Japanese visual culture and calligraphy. The project team responsible for the task (HdM and AUTH) will use the V4Design tool in order to find relevant visual material, which can be used in the design process. The scope of design includes furniture and exhibition partitions as well as signage and lighting fixtures. The creative team will use V4Design to browse for elements related to Japanese or Far East Asian culture and experiment with them. V4Design will be used to support the design process by making content (2D & 3D) available to the designer within his design authoring software (Rhinoceros3D).

PUC3: Design of virtual environments, related to TV series and VR video games architectural design, related to artworks, historic or stylistic elements

An audio-visual production company and broadcaster wants to create a VR video game based on the scenes and objects that exist inside a telenovela they produced. They will use the V4Design gaming authoring tool to build a gaming environment and gameplay for this telenovela. Using the tool they can remodel scenes and environments, and deploy a VR environment. The 3D models of interior elements will be extracted from the telenovela scenes and be provided to the video game creators to build a VR video game with the same assets, scenes and characters. Further 3D models bas on other art related content will be extracted from other visual cultural heritage sources to complement the scene and enhance the artistic features of the video game.

PUC4: Design of virtual environments, related to actual news for VR (re) living the date

Scenario 1, Gendarmenmarkt: DW (as a production company) wants to develop an interactive and immersive documentary using the existing footage they have from various news and locations. They have in-house software engineers who have a basic knowledge of the video game engine Unity and designers as well as dedicated journalistic storytellers. Selected parts of the archive will be imported to the V4Design platform and will be transformed to 3D and VR environments that will allow users to have an immersive experience, allowing them to relive a significant day in history.

Scenario 2, Bauhaus: The Bauhaus scenario extends the workflow of the Gendarmenmarkt scenario. From the additional material provided by shooting new video and 360° material, V4Design will be extended with a new range of archetypes, allowing to also integrate principles of “Bildnerische Gestaltungslehre.” DW editors will at first gather sets of Biedermeier images, models and textures through V4Design. They can search the database for objects or styles that resemble the provided samples. The content curator will then provide existing and new material on Bauhaus with exterior and interior shots to the 3D extraction processing engines of V4Design. The resulting 3D model can then be used as an environment for a VR experience to relive the origin of Bauhaus into modern times. DW’s engineers will then turn the building models and individual assets into several relive the date stories - from exterior perspectives and surroundings into interior scenes that showcase the progress of art and styles.

Figure 5: Gendarmenmarkt square

The main goal of the initial dataset is to provide content that can support these PUCs. In iteration 1 the keyword lists were used as main criterion to gauge the relevancy of data, while in iteration 2 queries were expanded to allow for content that didn't fit these keywords but were of a high enough quality to be reused.

2.4.2 **Content delivery process overview**

The **methods of collecting and delivering data and metadata** differ between content provider partners. Some content providers, like DW and EF, have APIs that allow for easy extraction of datasets that are customisable in granularity and complexity. Other data stores like those of AF and SLRS require manual curation work to extract content that is relevant to V4Design. The general steps in the content collection and provision process are outlined below.

1. Definition of content user requirements by user partners and content technical requirements by technical partners (Appendix A)
2. creation of keyword list (Appendix C) per PUC by technical partners
3. Extraction of content:

Data providers

a. EF and DW create queries based on keyword lists to call on the EF and DW APIs. Results are downloaded and sent to an FTP server
b. AF, SLRS, and DW use the keyword list to manually select content from their data. They download this data, create extra metadata like timestamps if needed, and send it to an FTP server
c. CERTH scrapes content from the DW Websites for PUC 3 and sends it to the FTP server

External sources

d. In the meantime, CERTH crawls and scrapes Website content like Twitter and Wikipedia to extract data based on the keyword list. The extracted data is

cleaned and sent to an FTP server. EF organises the collected data in the FTP server into a clear file tree (components of the dataset can be found in Appendix B)

4. EF creates documentation for their own data, other provider partners add to the documentation. The documentation is sent to the other consortium partners, together with access credentials to the FTP server.
5. Other consortium partners that have used the data fill in the evaluation form sent by EF (Appendix E).
6. The evaluation forms are taken into consideration in the definition of CURs and CTRs for the next iteration.

3 REQUIREMENTS AND EVALUATION

3.1 Content and Technical User Requirements

The requirements to successfully collect data that meets the needs of the V4Design technical partners as well as the user partners are split up into two categories: content user requirements and content technical requirements. The requirements were extracted from the HLURs (High Level user requirements), originally formulated in D7.2. The HLURs were split and listed per PUC, to cover all of the possible use cases that could occur in the use of the V4Design platform. These HLURs form the basis of the data collection requirements, which were then used to collect data for every use case. The Content User Requirements (CURs) and Content Technical Requirements (CTRs) per use case are listed in Appendix A. Their corresponding HLURs can be found in D7.1.

3.2 Evaluation for subsequent iterations

Throughout the first ten months of the work in Work Package 2, research results and feedback from other partners were taken as opportunities to adjust the current work process, capitalising on the strength of iterating over several versions of the initial dataset. The iterative agile process of WP2 and the CRISP-DM model command that we take feedback into account during the project, and not only use it for evaluation at the end of the project. More specifically, after the first iteration the user and technical requirements were changed. This subsection describes the adjustments to the tasks for which data would be gathered. The changes in gathered data can be seen in the rundown of the second iteration of the initial dataset in *4.3. Iteration 2: Qualitative phase* - overview of this dataset. Below are the most important content adjustments to the original work plan after getting feedback from iteration 1.

More focus on external data sources for 3D content

After an in-depth evaluation of the first iteration of the initial dataset, KUL gave the content provider partners extensive feedback on the data that was delivered for the 3D extraction and reconstruction tasks. Most of the video and image content that was delivered was deemed unsuitable for 3D extraction and reconstruction because they didn't meet the requirements for 3D extraction of 2D video content (see list of requirements in subsection 2.3). Since the 2D content from the first iteration of the dataset could not meet these requirements, we sought to create partnerships with external sources of 3D content that could be sourced for 3D segmentation tasks. We have first contacted Scantheworld and SketchFab (see subsection 4.1).

Pivot of tasks to pattern recognition and object segmentation

During the second iteration, the content provider partners of the V4Design consortium have tried to gather data that meets the requirements for new tasks that were formulated during the feedback phase of the first iteration, adding pattern recognition and object segmentation to the tasks of 3D extraction and segmentation. These tasks are still in scope for the V4Design project, and the data the content providing partners can extract meet the requirements for the execution of these tasks. Segmented 2D objects extracted from larger 2D images were identified by the user partners as useful objects to make available in the

V4Design platform. Identifying surfaces and textures that can be extracted from 2D content and placed over 3D objects as masks is seen as very valuable resources for video game creators, architects and other 3D designers.

Adjustments to video content

The main feedback on video data was given by both KUL and UPF. KUL advised on the usefulness of video data for 3D extraction and reconstruction, while UPF gave a more general evaluation of the usefulness of the objects in the video data for the V4Design platform, as well as the textual metadata that was delivered together with the video data. While UPF evaluated the video data from iteration 1 as a 4.5 on the Likert scale, the general feedback from KUL was that little to no video content was relevant or useful for 3D reconstruction.

Below is a list of the requirements for 3D extraction that couldn't be met by the historical cultural heritage videos possessed by the content provider partners:

- high-quality footage: most of the historical camera footage does not have a high resolution; although content provided by content provider partners that do have access to more recent camera footage have data that conforms to this requirement. Even then, the amount of suitable data with that resolution is too low to extract a high number of 3D objects from. Newer camera footage also has a higher statistical chance of still being and copyright, and would thus require licensing fees to be reused. In the worst case, the holder of the copyright of this footage doesn't allow any reuse at all.
- camera motion that allows to extract the features of the 3D space the camera is capturing. No 3D can be extracted from a static camera viewpoint, even if the shot is zooming, panning, or tilting.
- limited occlusion: when parts of an object are occluded in a shot, video footage can still show the occluded space later in the shot through movement of the camera or the object occluding the space. Issues arise when an object is largely occluded for most of the shot or parts that are occluded stay occluded for the duration of the shot.
- containing relevant objects for V4Design users: while a movie might in general be relevant to the potential users of the V4Design platform, this doesn't necessarily mean that design or architectural objects are shown in the clip are suitable for 3D reconstruction.

After discussing the issues listed above, several adjustments to the workflow of the creation of the second iteration of the initial dataset were made.

To collect data that can be used by KUL for 3D reconstruction and extraction, the following types of data were focused on in the second iteration of the initial dataset:

- 2D image data taken from different angles, extracted from external Websites like Flickr (*section* 5.2.3. *Flickr*).
- 2D image data from objects that are symmetrical in physical shape, extracted from Europeana (*section 4.1.1 Overview of this dataset*).
- 3D object data from external providers like Scantheworld and SketchFab (*section 5.1.1. Sketchfab and section 5.1.2 Scantheworld*).

- 2D video data for object detection and tagging, extracted from the consortium content providers
- specific drone footage from architectural landmarks and spaces, optimised for 3D reconstruction, extracted by Artfilms (*section 4.4.1 Overview of this dataset*).

Text

Textual data was mostly reviewed by UPF, who are responsible for the tasks in the consortium that deal with textual data. Every type of data by every content provider was evaluated on a Likert scale. Most of the evaluation was positive, with minor adjustments needed with regards to encoding, language fields, and unexpected characters. The biggest issue with the first iteration of textual data was the overall lack of large descriptive strings of text about architectural objects and designs. To accommodate this, in the second iteration of the initial dataset the Europeana Newspapers dataset was delivered to UPF. Europeana Newspapers contains over 11 million pages of international historical newspapers from all over Europe. It is a heterogeneous dataset spanning multiple languages and alphabet scripts, spread over the last three centuries. OCR has been performed on 6 million of these newspaper pages, which makes them fully searchable down to the article level. This immensely valuable data collection, first created by the European Library Project, has now been reused in the second iteration of the initial dataset for textual extraction in V4Design.

Images

Most of the feedback on 2D image data came from AUTH, who evaluated every dataset in the first iteration on a Likert Scale and provided feedback on their usefulness. This evaluation was highly useful for the tuning of keywords and search queries in the second iteration, which resulted in a much higher precision of returned items than in the first iteration. Most of the provided data was identified as mostly being useful for pattern and texture extraction, and object detection. In the second iteration of the dataset, more focus was laid on providing objects that could be used for these tasks.

CERTH also gave feedback on the first datasets that were delivered for aesthetic extraction. This dataset was a collection of about 17.000 paintings from the Europeana database, and their associated metadata. This data was highly relevant, but the biggest issue for aesthetic extraction tasks lay in the heterogeneity of the delivered data. With this heterogeneity of providing institutions, countries and languages it is hard to find a ground truth of styles and artistic movements for paintings. Developing a controlled vocabulary of styles and artistic movements from this highly heterogeneous data would warrant very resource-heavy data cleaning tasks, which is out of scope of the project. Instead, for the second iteration of the dataset a data collection of over 74.000 paintings and other artworks was extracted from Wikidata and Wikimedia commons, because they all contained metadata with controlled vocabularies for genre and/or artistic style.

Metadata

Usefulness and feature selection for metadata was mostly evaluated by CERTH, who are responsible for the ontological modelling in WP5 of the V4Design project. All metadata was deemed relevant, although work needed to be done on creating a more consistent set of features across dataset of different content providers. Lastly, to ease the ontological modelling process and to make data ready for Linked Open Data compliance, plans were

made to map other data models like EDM to SIMMO and deliver all data in SIMMO format for the final dataset.

4 DATA COLLECTION FROM CONSORTIUM PARTNERS

The following two sections, *4: Data collection from consortium partners* and *5: Data collection from external providers*, describe the actual contents of every iteration of the dataset. A list of the contents in every iteration of the dataset can be found in Appendix B. For every iteration a short description of the goals and requirements of that iteration are described, followed by a description of the objects in that iteration. Section 4 looks at content coming from the databases of consortium partners, and section 5 looks at content coming from social media, Websites, and external partners.

4.1 Iteration 1: Exploratory phase

4.1.1 Overview of this dataset

A list of components in this dataset can be found in Appendix B.1.

The goal of this dataset is to provide the consortium partners with a small selection of data that can be used as a starting point to:

- further define requirements
- become familiar with the data formats and types delivered
- testing technical workflows
- sparking discussions about and further developing the PUCs

Because the initial PUCs weren't finalised yet in D7.1, the actual subjects of the data were more fluid than in later iterations. In the second phase of this iteration, the PUC topics would be more strictly defined and therefore the data collection tasks gathered data of a higher quality and quantity. For this first phase, no keywords were defined, resulting in very differing subjects of data from the different content providers.

The requirements for this data collection task were as follows:

- collect at least 3 records of data that represent the data types and structure that the rest of the consortium can expect from future data delivery.
- deliver metadata that is as complete as possible, so a selection can be made and used in future data collection requirements.
- provide data in the most prevailing data format your database contains.
- provide data on different subjects, keeping the target user group of V4Design in mind.
- The data should be freely reusable for V4Design consortium members, not necessarily for the general public.

This work was conducted in M1-M3 of the project. With the technical architecture not yet set up for receiving data from content providers, a private wiki environment was created in which URLs and URIs could be deposited that linked to data stores where the collected content could be viewed and/or downloaded. This solution balanced the speed and ease of use needed in the first stages of this project with the need for a robust and reliable store of data that content providers could use to deliver their product.

Europeana delivered 8 different small data collections gathered from different simple API queries. The queries were often single terms without any more filters except the specification that the objects should be freely reusable. The CHOs all contained image files in

the most often encountered image formats (.jpeg and .png). The metadata for these CHOs was delivered in JSON, with the profile query parameter for this metadata set to rich, so as to catch the largest number of metadata fields possible. Examples of data from these data collections are photographs of Hans Scharoun's architecture, photographs of the Berlin wall, historical art objects from what is currently the Hong Kong area, sketches from the layout of the Versailles palace gardens, and blueprints of the Notre Dame de Dijon.

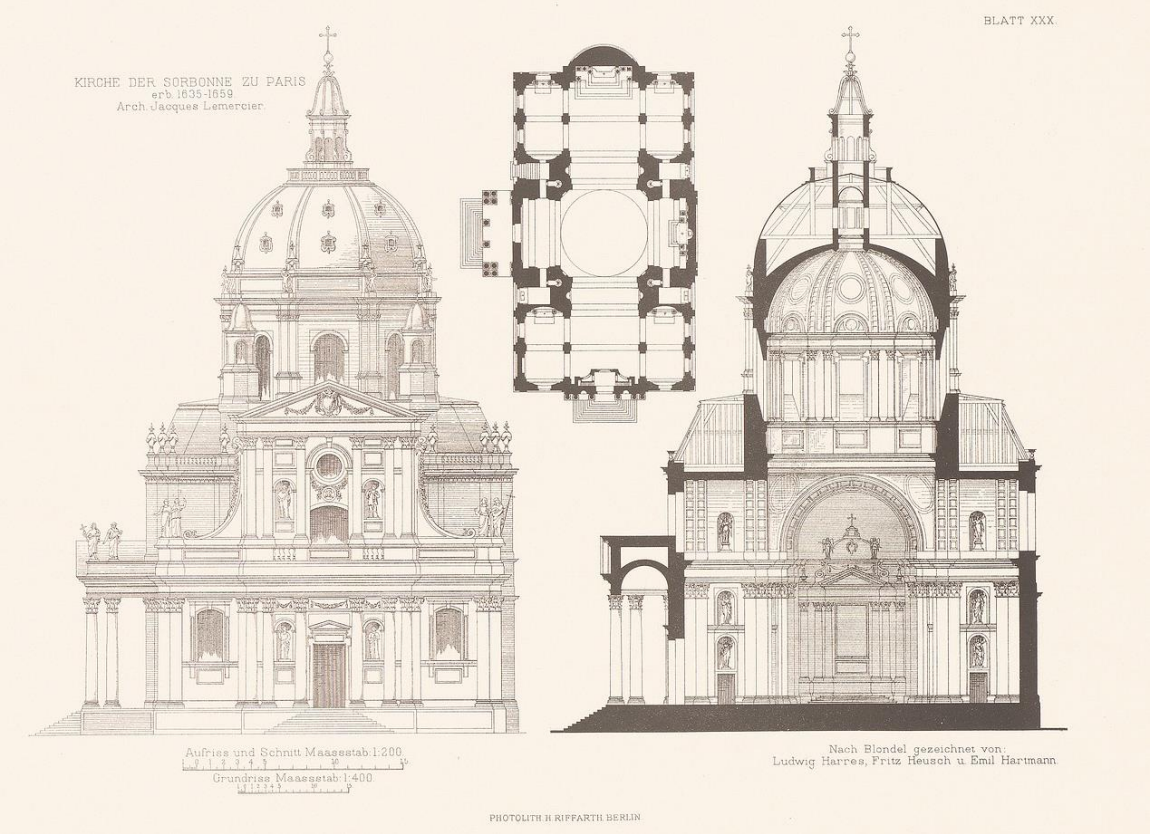

Figure 6: "Kirche der Sorbonne zu Paris"- Architekturmuseum der Technischen Universität Berlin in der Universitätsbibliothek - Germany - CC BY-NC-SA

ArtFilms delivered 15 full length video and 1266 images supporting the need for architectural objects; buildings, cityscapes, parts of buildings. This initial dataset included documentaries about Bauhaus buildings and some well-known architects' work, also featured landscapes and cityscapes from China, Indonesia, Japan, Greece and the Mediterranean as well as from Australia. The photographs provided featured buildings, cityscapes, textures, ornaments and objects from Europe, China, Japan and some other part of the world.

The entire Artfilms collection of 1700 films was made available for partners to browse and look for initial ideas and topic focusing on architecture and design. A selection of films was identified and links were provided from Artfilms' dropbox to these films featuring objects, cityscapes and buildings.

Figure 7: (left) Warehouse building: windows and wall. (right) Chinese ornament wall

Deutsche Welle focused at the beginning of the project on screening a) what kind of DW content is relevant for the project, and b) evaluating how the content can be made available. For PUC3 and PUC4 different kinds of content are required: PUC 3 requires access to the audio-visual and textual content of Nico's Weg. This content is available online, but can't be accessed via API. In close collaboration and consultation with DW's system architect for API's it became apparent that the status quo might change towards the end of the project but for now another solution was needed in order to not be dependent on internal processes. All episodes of Nico`s Weg are accessible through YouTube, the accompanying exercises via DW's language learning platform. For PUC 3 as well as for PUC 1 & PUC 2 we screened DW's photo and video material published on the media centre, and created a list of presumably relevant content. This material is accessible via DW's regular media API.

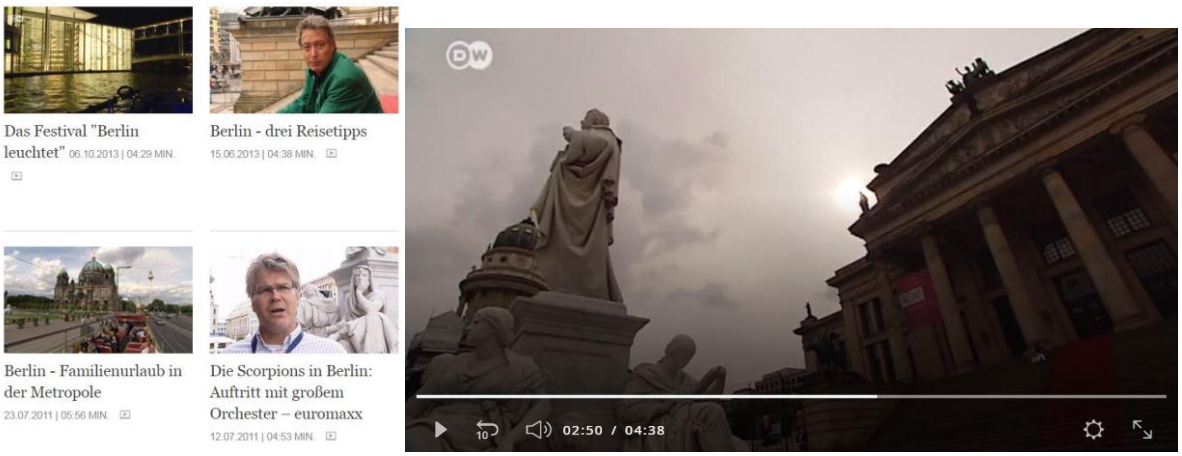

Figure 8: Screenshot of DW Media Center, which is accessible via API, and screenshot of an actual DW video listed as a result of a search query showing the Gendarmenmarkt.

Solaris Multimedia AB delivered three full length videos in MP4 format including Kochuu, Great Expectations and Microtopia, to meet the project request for content about Japanese and Chinese art and architecture, and modernist 20th century architecture. Examples of data provided are video portions from the film "Microtopia" and video portions from the film "Kochuu - Japanese Architecture, Influence and origin".

Figure 9: (left) Hiunkaku Pavilion (still shot of Japanese Architecture). (right) Habitat 67 (traveling shot of modernist architecture).

4.2 Iteration 1: Quantitative phase

4.2.1 Overview of this dataset

A list of the 3 components of this dataset can be found in Appendix B.2. The first phase of this iteration and the first months of the V4Design project resulted in clearly defined PUCs, a list of requirements from both users and technical partners, and an initial technical architecture solution for providing data flow from content providers to the other consortium members. For the second phase of the first iteration, data was uploaded to the FTP server. Documentation for the data moved from the internal wiki to a dedicated documentation file. The goal of this phase of the dataset was to:

- provide a large quantity of data that could be used for simulation examples and workflow testing by technical partners
- extract as much data using the predefined keywords from the PUCs to create a dataset from different content providers and in different formats but about the same subjects.
- adhere to the initial requirements set out by the users and technical partners and gather feedback on the quality of the gathered data, amendments to the requirements, and feedback about the feasibility of using this data in the workflow of the scientific and technical objectives.

At the creation of the datasets in this phase, which together with the datasets of phase 1 are considered the complete dataset of the first iteration, the Data Management Plan (DMP) was updated to reflect the creation of this dataset. The next update of the Data management and self-assessment plan (Version 2, delivered in D1.3.) will reflect these changes. In accordance with the guideline on data management in Horizon 2020, the following points are processed in the Data Management Plan: dataset reference and name, dataset description, standards and metadata, data sharing and archiving and preservation (including storage and backup).

Europeana provided both visual (images) and textual (metadata) data for this iteration of the initial dataset. The CHOs and the metadata that accompanies those objects have all been extracted from the Europeana databases using the harvesting script written specially for V4Design. The script is written in Python 3, uses a few non-standard modules, and calls on the Europeana Search API. Metadata is stored in JSON format, in a single JSON file. The metadata was harvested with the profile field = rich, so it includes the maximum amount of metadata that can be harvested. All metadata from Europeana is available under a public

domain CC0 license. All objects have been harvested with the reusability field = open. This iteration of the dataset contains approximately 23.400 images and 23.400 JSON objects with metadata of those images. The queries that extracted this data strictly followed the keywords of the PUCs, which can be found in Appendix 3. Apart from providing large data collections on the subject of the PUCs, Europeana also started providing data for more specific tasks in the V4Design consortium. More specifically, an individual dataset of approx. 14K CHOs was delivered for aesthetic extraction (T3.5).

Artfilms provided 29 videos and 1266 photographic stills. The video material was transcoded into Mp4 files and was uploaded to V4Design's FTP platform organized in folders relevant to the users' requirements. A file with timestamps was also provided to help the users to find the clips which might be relevant to the project. A representative example of the videos might be the film "The Treasures of Ancient Hellas: The Parthenon" which includes maps, drawings and 3D reconstructions of the various facets of the Parthenon of the 5th, 6th century.

Deutsche Welle delivered video and textual items for PUC3 and PUC4. Next to the YouTube videos and exercises for PUC3, we added the screenplay of the language learning videos for level A1 and B2 as PDF as well as transcript examples of the videos as .doc files (level A1, cluster 13, number 1 to 4). After screening DW's Media Center, ergo online accessible content, DW searched for further video content of the Gendarmenmarkt in our archives in order to provide content for PUC4. Deutsche Welle has video material since the 90s, but not all of it has been digitized or made accessible online. Thus, we consulted the *Fernsehdatenbank* (Fesad; television database), where we find all archived video items (some of the news items show video footage provided by Germany's regional public-service broadcasters). Finding adequate footage in Fesad is highly depending on the quality of tagging, which ranged from very good to poor. However, seven videos could be identified as potentially relevant for the project (no preview existed), thus, DW collaborated with the DW's Archives to request a digital version of those videos, of which DW would ultimately deliver two (.mxf file) to the V4Design database. Furthermore, DW shot additional footage of the Gendarmenmarkt and added it to the database.

Solaris delivered three full length videos including Kochuu, Great Expectations and Microtopia, to meet the project request for content about Japanese and Chinese art and architecture, and modernist 20th century architecture. The films were uploaded to the V4Design ftp in mp4 file format in March 2018.

In order to make the films more readily searchable, timecode-named stills of keyframes from each cut from the films were delivered for the documentaries Kochuu and Great Expectations, including Excel-files with timecodes referring to these timecode-named stills of keyframes from the documentaries Kochuu and Great Expectations. Examples of data provided are keyframes from the film "Kochuu" and keyframes from the film "Kochuu - Japanese Architecture, Influence and origin"

Figure 10: (left) Nakagin Capsule Tower (Japanese Architecture by Kisho Kurokawa) (right) Unité d'Habitation (facade, modernist architecture by Le Corbusier).

CERTH exploited the initial crawling and scraping module to collect textual information concerning the Deutsche Welle Nico's Weg courses[7] for learning German. 440 webpages were crawled and scraped from this domain. More specifically, CERTH crawled the domain of the courses intended for beginners learning the A1 level of the language. Pages presenting grammar, vocabulary and culture/society details were scraped and integrated into this dataset. An example of such webpage is the one that describes the formal and informal way of greeting somebody and its extracted content is shown in Figure 11.

Saying goodbye: informal
Tschüss, Martina! – Mach's gut!
Bis bald! – Tschüss!
The informal expressions for saying goodbye are used with family, friends and acquaintances, and often with colleagues.
Friends, acquaintances and family members are addressed by their first names.

Saying goodbye: formal
Auf Wiedersehen, Herr Tillmanns! – Auf Wiedersehen!
The formal expression for saying goodbye is used with unfamiliar adults, in business situations, with officials, and when shopping.
Men are addressed formally as Herr followed by their surname, women as Frau followed by their surname.
If you are unsure whether to be formal or informal, use the formal expression for saying goodbye.

Grammatical terms in German:
informell: Informal language is used when talking to family and friends.

Figure 11: Scraped text from https://learngerman.dw.com/en/informal-and-formal-1/l-37250531/gr-38310488.

[7] https://learngerman.dw.com/en/beginners/c-36519789

4.3 Iteration 2: Qualitative phase

4.3.1 Overview of this dataset

A list of the components of this dataset can be found in Appendix B.3. The second iteration of the initial dataset is also the definitive iteration, combining with the earlier iteration to form the completed initial dataset finalised with D2.1. This iteration breaks away from the keywords and PUCs set forth in D7.1 in favour of creating discrete data collections tailored to the needs of individual technical or scientific tasks in the V4Design project. With the completion of the first simulation examples, and feedback on the first iteration leading to amended technical and user requirements, the content providers now had enough information to deliver datasets that can be used for various tasks within the V4Design project. The goal of this second iteration was to:

- Take in the feedback from iteration 1
- Expand the amount of objects that are usable by the consortium
- Increase the quality of usable objects by the consortium
- Increase the amount of external sources from which content can be sourced

Below is a rundown of data delivered, per Work Package.

WP3: Visual and Textual content analysis

Europeana provided UPF with a data dump of the Newspapers dataset handed down to them from The European Library project. This dataset was collected by The European Library over the span of three years, from 2012 to 2015, in partnership with a high amount of national libraries of several different European countries. This culminated in over 10 million digitised and historical newspaper pages with OCR full text being collected and provided freely in one place, available for research, education, and curiosity. After the end of the European Library project, Europeana took over the legacy dataset and started work on integrating that data into their Collections platform. The launch of the Europeana Newspapers collection is scheduled for December 2018. Europeana did provide a provisory data dump of about 6.6 million digitised newspaper pages, originating from 592 different newspaper titles. This massive textual dataset lies entirely in the Public Domain and can thus be studied and reused in any way UPF sees fit in the pursuit of their technical and scientific objectives. When the Europeana Newspapers collection gets finalised and released to the general public, an update of this dataset will be delivered to UPF for T3.1.

A dataset to support the aesthetic extraction work performed by CERTH was created by Europeana together with support from CERTH. This dataset is the Sum of all Painting data collection extracted from the Wikidata and Wikimedia LOD database.

WP4: 3D model extraction from 2D visual content

Europeana created a dataset of 2D images with the goal of extracting 3D out of images containing symmetrical objects, such as vases, jars, and amphorae. This dataset was delivered to accommodate a deviation in the research goals of KUL, towards the exploration of the possibility of extracting 3D out of a single 2D image, or multiple 2D images, if there is prior knowledge that the object is symmetrical in shape. When the relevant models have been run on this dataset KUL will be able to provide feedback about the feasibility and

technological readiness of this objective, which will in turn inform if more datasets of a similar type will be collected for inclusion in the Final V4Design dataset.

ArtFilms collected several videos of drone footage for KUL following their feedback requesting more high-definition footage from non-static camera angles. This drone footage should return a higher quality and quantity of 3D models using KUL's 3D extraction tool for video footage.

5 DATA COLLECTION FROM EXTERNAL PROVIDERS

Next to collecting and analysing content from the content providers, the V4Design project also envisioned to crawl textual and visual content from external sources to complement the data that V4Design collects from its consortium partners. Most of the data collection for this task was performed by CERTH in ***"T2.1: Web Crawling and retrieval of textual and multimedia data".*** After the need for adjustments in the content provision were identified by the feedback, more resources were put into collecting data from external providers. More specifically, it has become necessary to look at other providers of freely reusable 3D content. The rationale behind this necessity has been explained in section 2.3.

The first section describes the collection of 3D resources from external partners, mainly conducted by EF. The second section is a rundown of the data collection tasks performed in T2.1, as well as the data collection tasks performed in support of and in coordination with ***"T3.1: Compilation and study of texts relevant to visual data".***

It is important to highlight the legal and ethical implications of using data sources that are external to the data partners of the V4Design consortium. All data partners in the consortium consent actively to provide their data for research and reuse in the V4Design project and on the V4Design platform. Furthermore, all data partners have extensive influence in exactly what data, how much data, and in which format their data is queried, extracted, and provided to the consortium. This high level of influence ensures that there are no legal or ethical boundaries that are crossed in the use and re-use of data from content provider partners in the V4Design consortium. Lastly, all content provider partners are aware of the extent of the use and re-use of their data in the project, and they ensure that all of the data they provide is cleared for free re-use.

With data from external partners however, it becomes increasingly important to safeguard that the same legal and ethical boundaries set out by the content provider partners aren't overstepped as well. The difference here lays in the fact that external data providers often do not have an active say or influence in how the V4Design project utilises their data. If data is available on the internet, this does not automatically mean it can be used and re-used without issue. When crawling or scraping data from social media platforms, a high percentage of the data that is available might not be cleared for re-use. For instance, at the time of writing 99.8% of images hosted on the image Website Flickr are not freely reusable and can thus not be crawled for use in the V4Design project.[8] It's imperative to parse the data that can be reused from that that cannot be reused before crawling or scraping, and providing ample documentation and rationale of why and how this parsing was conducted.

In the case of the sources of 3D data (section 4.1.), direct contact was made with the providers of this external data to first ask them for input, requesting clearance to crawl their data, and clearly communicating what data would be extracted and how it would be used in the project. This gave the external 3D data providers the opportunity to give input into the crawling and scraping process, and advise on how to best extract and reuse their data. In the

[8] on 18/09/2018, there were 44.6 billion images hosted on Flickr, of which 57.8 million either carried a CC BY, CC BY-SA, PDM, or CC0 license. https://www.flickr.com/creativecommons & https://www.flickr.com/photos/

case of scraping and crawling data from social media platforms and other aggregation platforms such as Wikipedia, a desk study was first performed to ensure the workflow of these tasks stayed within legal and ethical boundaries. More information on this can be found in section *4.2: data collection from online Web resources*, and section 5: *Legal and Ethical requirements.*

Technically, three broad techniques were used to extract data from data stores not provided by content provider partners. The first we refer to as content ***crawling*** which is the process of automatically collecting a list of URLs from target domains, using some seed webpages as entry points. The second technique is referred to in this document as content ***scraping***, and is a robust way of extracting data from the Web. It is used when no public APIs are available for a Web platform or service, or if the APIs don't give access to the content that needs to be extracted. With Web scraping, the actual HTML content that is served to the user when it connects to a Website is downloaded, often parsing the HTML tags and only extracting the content from relevant tags, such as images, metadata, or hyperlinks. Last, in the cases where a Website provides an API for extracting their data, the work done in crawling and scraping can be replaced by performing ***search*** on the API. Search is the act of connecting to the API endpoint of an external source and querying it multiple times to extract data from the backend of that source. It is most of the time a more effective way of retrieving information from a Web domain than employing crawling and scraping techniques.

5.1 3D data collection from external sources

5.1.1 Sketchfab

One of the top providers of online 3D artefacts is Sketchfab, allowing users to upload an infinite number of 3D models for free. A certain portion of those models are licensed under a creative commons license. Getting access to this vast store of user-generated 3D content for re-use in V4Design and for integration in the final V4Design platform is a great step forward to ensure that the users of the final V4Design platform can browse through a vast array of 3D models, and find the content that suits their re-use needs. SketchFab was contacted after feedback was given on iteration 1 of the dataset to ask if they would be interested in providing their 3D models as content for the V4Design technical and scientific objectives and for possible inclusion in the final platform. SketchFab reacted positively to this request, providing access to their API for easy download and reuse of their models, and showing interest in further talks that could lead to the inclusion of their models in the final V4Design platform. Developing a further liaison with Sketchfab is an important part of the future work that will be conducted between D2.1 and D2.3.

5.1.2 Scantheworld

Scantheworld is a 3D scanning initiative focused on providing high-quality 3D-printable scans of cultural heritage artefacts. It is a crowd sourced public effort to make cultural heritage more accessible with 3D scans. Scantheworld was also identified as a prime source of 3D cultural heritage objects that would be ideal to contribute to the scientific objectives of T4.4 *Enhanced 3D model extraction.* After fruitful discussions with Scantheworld to inform them of the project, they were very interested in supporting the project with their expertise and their data stores. EF and KUL met with Scantheworld to discuss possibilities of collaboration, and received a very positive response. Scantheworld provided their APIs for the extraction of

their 2800+ 3D models, of which almost all are freely reusable (>99%). All these models can be used by the V4Design consortium for research purposes, and Scantheworld is open to discussion about inclusion of these models in the final V4Design platform. The possibility of creating a two-way technical pipeline between The V4Design platform and the Scantheworld one was also discussed, opening up the opportunity for re-users of Scantheworld objects to re-upload their remixes of the original content to both the V4Design and the Scantheworld platforms.

5.2 Data collection from online Web resources

The goal of collecting information from online Web resources for this reporting period is to inspect various domains and check their data quality and suitability for supporting the other V4Design tasks (e.g. 3D reconstruction). Regardless of the methods and the performance of the crawling and scraping component that automates this data gathering procedure, an important factor that affects the calibre of its final output is the resources and the queries we choose to feed as input. Thus, experimenting with some well-known Web domains is a valuable step that must be seriously taken into consideration.

5.2.1 Wikipedia

A widespread source of information for any kind of topic is Wikipedia. We are interested in both the textual and visual content on Wikipedia, as text is going to be given as input to the V4Design text analysis tasks, for example the named entity identification, whereas the visual content is need for tasks such as the aesthetics extraction and the 3D reconstruction. An appropriate entity that forms an example for most tasks of this project could be a castle. Therefore, we chose to scrape information from 311 Wikipedia Web pages describing castles for the first iteration of the dataset. The extracted text includes the main content of the webpage as well as metadata placed in the infobox that exists at the majority of the Wikipedia pages. Images existing in the webpages are also collected along with their captions. Due to the fact that additional resources from Wikipedia correlated with PUC4 were needed to be evaluated, three more webpages that were not relevant to castles were scraped and added into the dataset. A total of 661 images were saved from the entire URLs set. An example page that was crawled is the Neue Kirche (New Church) which is situated in Berlin. The image existing in its Wikipedia page infobox is illustrated in Figure 12. All the extracted data were stored in a MongoDB database; two collections were created for the textual and visual content respectively. Their delivery was made by extracting a JSON formatted text out of each collection.

Figure 12: The New Church on Gendarmenmarkt, seen from north

5.2.2 **Twitter**

In the framework of the T2.1, CERTH plans to extract information from social media resources apart from the typical websites. Scraping text from Twitter could be useful for text analysis tasks in the same way as the texts in Wikipedia. In this case, we are interested in content about Cultural Heritage. To this end, we leveraged the Twitter API to search tweets posted by a limited set of users. 21 users were selected for the search which included: a) Twitter bots that continuously tweet out objects from digital archives like the Metropolitan Museum, b) official Twitter accounts of Cultural Heritage institutions, c) other accounts related to Cultural Heritage not falling into the previous two categories. The final outcome consists of texts from about 40 thousand tweets. All contents are stored in MongoDB and they were delivered in JSON format. A typical instance of such Twitter post is located at https://twitter.com/MuseumBot/status/892811000994557953. That post was published from a bot tweeting images from the Metropolitan Museum of Art.

5.2.3 **Flickr**

Flickr is a popular image and video hosting application. Any registered user can share and organize its photos/videos using this platform. It contains one of the largest multimedia databases available in the Web. As a result, we can make use of it to form our V4Design collections. For the needs of the project tasks, we focused on searching and collecting images from famous buildings as well as objects such as cars. We made requests to the Flickr Search API using 13 queries related to the aforementioned categories as input. Example queries are the Eiffel tower, the Delphi temple and the Volkswagen Beetle. By setting a maximum number of results per query to one thousand and by taking into consideration the licensing restrictions for reusing the content, we formed an image dataset of 6209 items. The main criterion this collection is to be tested is its appropriateness for the 3D reconstruction task. Most of the details accompanying each returned image are stored in MongoDB and they were delivered in JSON format. An example item of the constructed Flickr dataset is shown at Figure 13.

Figure 13: A picture of Eiffel tower existing in the Flickr Website

5.2.4 **Future datasets**

As for the next steps towards collecting data from the Web, the first one is to define more Web resources and use (or upgrade if needed) the already developed crawling and scraping infrastructure to create more datasets. Furthermore, more improvements are to be introduced in order to refine the final output. Optimizations can be done either by improving the HTML rules based on which webpages are scraped or by requesting and saving more information (e.g. metadata) while searching from the available APIs. For both updates, more significantly for the latter one as APIs already provide a variety of options for the retrieval output, the fields that are indeed useful for the tasks that are dependent on these data shall be investigated and determined.

6 LEGAL AND ETHICAL REQUIREMENTS

According to the DoA, D2.1 will contain a comprehensive description of the legal and ethical requirements in relation to the use cases. The DoA distinguishes between the two most relevant legal aspects (1) Privacy and data protection and (2) Copyright and other intellectual property rights. The following text will reflect this distinction. Its aim is less a comprehensive legal analysis of data protection or copyright law in general, but to create an overview and a form of legal manual for the successful implementation and usage of the V4Design tool.

The background for this legal manual is D1.2, the "First version of the Data management and self-assessment plan". In section 2 of this deliverable specifically, the consortium has enlisted and described a number of datasets that will be generated and used within the different V4Design WPs. In detail, D1.2 mentions

- a WP1 dataset that contains contact information for project partners and advisory board members;
- a number of WP2 datasets deriving from multimedia data crawling and containing files of images, texts, videos and other content from various sources (consortium partners as well as Web sources);
- a number of WP3 datasets consisting of texts in the different V4Design languages with their associated summaries, annotations or other information that are relevant to the end users;
- a number of WP4 datasets containing data that is created in the process of 3D modelling;
- a number of WP5 datasets containing the semantic representation of the annotations that are generated by the various V4Design modules as well as additional linked data;
- a WP6 dataset consisting of the requirements and component descriptions related to the V4Design architecture - including data about their development;
- a WP7 dataset that is generated by the users of the V4Design platform such as users' personal information, detailed logs of user actions or information about user devices;
- a number of WP8 datasets containing all information (such as contact details) that is relevant for project dissemination as well as for determining key dissemination indicators.

It is obvious on first sight, that these datasets include information that falls under data protection law (e.g. contact details) as well as content that is protected by intellectual property rights (e.g. images or texts). Some of this data, on the other hand, is legally not relevant at all. In the two following subsections, we will briefly describe the foundations of data protection law (subsection 5.1) and copyright law (subsection 5.2) that are relevant for V4Design. Subsection 5.3 will briefly assess the relevance of other IPR law (e.g. design law) and Subsection 5.4 will describe the legal consequences for each of the datasets that have been mentioned in D1.2 and suggest protocols in order to comply with these legal requirements.

The legal analysis will be based on European law as well as on the legislation of those countries that are represented in the consortium. Additional legal analysis is necessary If data or content is used that is governed by other jurisdictions (e.g. France or Italy).

6.1 Privacy and Data protection

On December 7th, 2000, The European Parliament, the Council and the Commission solemnly proclaimed the Charter of Fundamental Rights of the European Union. Article 8(1) of this charter provides that

- everyone has the right to the protection of personal data concerning him or her, and that
- such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. It adds that everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.

In 2016 and after years of controversial discussion, the EU passed a *General Data Protection Regulation* (GDPR) in order to further specify these fundamental rights (Regulation (EU) 2016/679). The regulation has come into effect on May 25th, 2018 and is since then directly applicable in all member states. The GDPR regulates that any processing of personal data can only be based on the data subject's personal consent or a conclusive number of other reasons laid down by law. Furthermore, the GDPR introduces strict compliance rules for data controllers as well as severe sanctions in case of any violation of the regulation itself.

6.1.1 Personal data

Article 4 (1) GDPR defines "personal data" as *any information relating to an identified or identifiable natural person* (the 'data subject'). An identifiable natural person in this sense *is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*. It is discussed controversially whether a person needs to be identifiable by the specific data controller in question (who could have restricted means and sources) or whether it is sufficient that this person is *theoretically* identifiable by someone. Data protection authorities (e.g. in Germany) tend to lean towards the latter view. A ruling by the European Court of Justice from 2016 also tends to a wide understanding of personal data by declaring *dynamic IP addresses* as personal data (Judgement of the Court of October 19th, 2016 in case C‑582/14). Therefore, and in order to be on the safe side, the consortium should consider any piece of information that could somehow be related to a specific data subject as personal data.

For V4Design, this means that all information that clearly identifies a specific data subject (names, contact details etc.) is personal data. But even anonymised or pseudonymised information or any piece of information that could theoretically (e.g. by big data analysis) be related to a specific data subject should be considered as personal data.

6.1.2 Data processing

According to Article 4 (2) GDPR, "processing" means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure

or destruction. Given this wide definition, any dealing with personal data by the consortium should be considered as "data processing" in the light of the GDPR.

6.1.3 Data processing on the basis of personal consent

The most straight-forward way to obtain legal permission for the processing of personal data, is to ask for the data subject's personal consent. According to Article 6 (1a) GDPR,

processing shall be lawful […] if and to the extent that […] the data subject has given consent to the processing of his or her personal data for one or more specific purposes.

Article 7 GDPR further specifies the conditions for consent by especially requiring that

the request for consent shall be presented in a manner which is clearly distinguishable from […] other matters, in an intelligible and easily accessible form, using clear and plain language.

Prior to giving consent, the data subject needs to be informed that he or she has the right to withdraw his or her consent at any time without giving any specific reason. It needs to be as easy to withdraw as to give consent. However, any withdrawal of consent does not affect the lawfulness of processing based on consent before its withdrawal.

With regard to V4Design, this means that in all cases that we ask for the personal consent of data subjects, we will do this by explaining in clear and plain language for which specific purposes we intend to process the data. A template for a consent form can be found as Appendix F to this deliverable.

6.1.4 Data processing on the basis of other reasons laid down by law

Although personal consent is the most straight-forward legal ground for data processing, it is often not the best choice or even a manageable option. First of all, the right to withdraw consent at any time causes risks as this withdrawal terminates the lawfulness of data processing for the future; a consequence that could be severe when, for instance, successful scientific research is based on the further availability of this data. Furthermore, the specific data subject(s) might not be reachable or quite simply not prepared to give consent. And finally, especially with regard to big data analysis, it might not be manageable to address all affected data subjects.

In order to facilitate lawful data processing in at least some of these cases, Article 6 GDPR provides a *conclusive* enumeration of additional legal grounds. One example are data processing activities that are necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract (Article 6 (1b) GDPR). Another example is data processing that is necessary to comply with legal obligations (Article 6 (1c) GDPR).

A further legal ground that could be relevant regarding data processing within the V4Design project is Article 6 (1f) GDPR. Here, the regulation provides that the processing of personal data is lawful if it is

necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

This means that, as a first step, we need to assess whether the interests of the V4Design consortium can be considered as "legitimate interests" in the sense of Article 6 GDPR. In a second step, we need to analyse whether these legitimate interests might be overridden by the data subject's fundamental rights and freedoms. In 6.1.7, we will determine whether the V4Design goals establish a legitimate interest for the processing of personal data.

6.1.5 **Fundamental principles of data processing**

Even if data processing can be based on personal consent or other reasons laid down by law, the data controller still needs to strictly follow a number of fundamental principles regarding the treatment of this data:

- *Purpose limitation*: Personal data need to be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes (Article 5 (1b) GDPR).
- *Data minimisation*: The processing of personal data needs to be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (Article 5 (1c) GDPR).
- *Accuracy*: The processing of personal data needs to be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay (Article 5 (1d) GDPR).
- *Storage limitation*: Personal data need to be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed. However, personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject (Article 5 (1e) GDPR).
- *Integrity and confidentiality*: Personal data need to be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures (Article 5 (1f) GDPR).

6.1.6 **Data processing Compliance rules and possible sanctions**

The GDPR also provides a number of compliance rules. Their purpose is to establish management structures and specific protocols within the data controller's organisation that ensure that the processing of data follows the GDPR's provisions. The most relevant compliance rules are:

- *Privacy by design*: The data controller is obliged to implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the GDPR's requirements. These measures need to take into account the state of the art, the cost of implementation and the nature, scope, context and purposes of

data processing as well as the risks for the respective data subject (Article 25 (1) GDPR).

- *Privacy by default*: The data controller is obliged to implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility (Article 25 (2) GDPR).
- *Records of processing activities*: The data controller is obliged to maintain a record of processing activities under its responsibility. That record is supposed to contain:
  - the *name* and contact details of the controller and, where applicable, the joint controller, the controller's representative and the data protection officer;
  - the purposes of the *processing*;
  - a description of the categories of data subjects and of the categories of personal data;
  - the categories of recipients to whom the personal data have been or will be disclosed including recipients in third countries or international organisations;
  - where applicable, transfers of personal data to a third country or an international organisation;
  - where possible, the envisaged time limits for erasure of the different categories of data;
  - where possible, a general description of the technical and organisational security measures that are taken in order to safeguard personal data. (Article 30 GDPR)
- *Data protection impact assessment*: Where a type of processing in particular using new technologies is likely to result in a high risk to the rights and freedoms of natural persons, the data controller is obliged to carry out an assessment of the impact of the envisaged processing operations on the protection of personal data (Article 35 GDPR).
- *Data protection officer*: Data controllers that regularly process personal data on a larger scale are obliged to appoint a data protection officer whose main tasks are to advise the data controller about data protection, to monitor compliance with the GDPR and to serve as contact point for the supervisory authorities (Articles 37, 38 GDPR). In Germany, for instance, any organisation with ten or more employees in charge of processing personal data, is obliged to point a data protection officer (§ 38 Federal Data Protection Act).
- *Administrative fines*: These compliance rules are accompanied by the possibility of severe sanctions in case of any violation of the GDPR. These administrative fines shall in each individual case be effective, proportionate and dissuasive. They can lead up to 20 Mio EUR or up to 4 % of the total worldwide annual turnover of the preceding financial year, whichever is higher (Article 83 GDPR).

6.1.7 **Legal framework for data processing within V4Design**

After having described the GDPR's general provisions, we will have a closer look at the legal framework for data processing within the V4Design project.

As pointed out before, any piece of information that could theoretically be related to an individual person, needs to be considered as "personal data". And any collecting, scraping, storing, analysing, clustering and sharing of this information is "data processing" that needs to follow the rules and principles of the GDPR.

The question is whether this processing of personal data is lawful or not. In all those cases where the consortium targets specific individuals (consortium members, advisory board members, end users etc.), the processing of their personal data could and should be based on personal consent. This means that they should be asked to sign a consent form (Appendix F) that clearly describes the scope and the purpose of the processing.

Other activities within the consortium that include the processing of personal data can most probably not be based on personal consent. This especially refers to data crawling that leads to the collection, storing and processing of files of images, texts, videos and other content from various sources. In these cases, it will not be possible and/or manageable to contact all individuals whose personal data are involved. However, the processing of personal data might be justified by the respective data controller's *legitimate interests*. As elaborated in section 6.1.4., the processing of personal data can be based on legitimate interest as long as the interests or fundamental rights and freedoms of the data subject do not override.

In the following, we will analyse the goals of the V4Design project (including its four pilot use cases) in order to assess whether they establish legitimate interest according to the GDPR:

Research

V4Design is a Research & Innovation Action so that it is an obvious first step to examine whether scientific research activities can establish legitimate interest that justifies the processing of personal data. In fact, the GDPR privileges scientific research in a number of sections. For instance, data processing for archiving purposes in the public interest, *scientific* or historical *research purposes* or statistical purposes is generally not considered to be incompatible with the initial purposes (Article 5 (1b) GDPR). This means that scientific research remains possible even if it was not covered by the initial purpose and even if the data subject had not been informed about this option beforehand. It is also not required to subsequently inform the data subject about the use of his or her data if the provision of such information proves impossible or would involve a disproportionate effort (Article 14 (5b) GDPR). Furthermore, this personal data may be stored for longer periods as long as the data will be processed solely for [...] scientific [...] research purposes (Article 5 (1e) GDPR).

On the other hand, these regulatory privileges for scientific research correspond with specific safeguards that researchers need to observe. Especially Article 89 GDPR contains a number of safeguards and derogations relating to data processing for scientific research. Its goal is to balance out the fundamental rights to privacy and data protection on one hand and the freedom of scientific research on the other. Altogether, Article 89 and the various other provisions in the GDPR lead to the following principles that need to be observed when the processing of personal data is justified by scientific research (see *Däubler/Wedde/Weichert/Sommer, EU-Datenschutz-Grundverordnung und BDSG-neu, 2018, DSGVO Art. 89, recital 32*):

- The processing of personal data for scientific research should be carried out on the basis of *anonymised datasets.* If the research purpose does not allow for

anonymsation, the datasets should at least by pseudonymised. The risk of *re-identification* should be minimised as far as possible.

- The processing of personal data for scientific research is only permitted if the responsible researchers establish organisational and technical measures that safeguard a maximum of integrity, confidentiality, transparency, availability and resilience of processing systems and services.
- The processing of personal data for scientific research *without the data subject's consent* is only permitted if it is not overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data.
- Researchers are only entitled *to publish* personal data on the basis of prior personal consent.

As a result, research activities within V4Design may establish a legitimate interest in the processing of personal data according to Article 6 (1f) GDPR as long as the aforementioned safeguards and limitations are observed. However, any data controller that justifies its data processing activities with scientific research needs to take into account that the GDPR grants member states a considerable degree of freedom and flexibility to develop their own regulatory framework. According to Article 85 GDPR, member states shall by law reconcile the right to the protection of personal data pursuant to the regulation with the right to freedom of expression and information, including processing for journalistic purposes and the purposes of *academic*, artistic or literary expression. Each V4Design consortium partner who intends to base data processing on scientific research as a legitimate interest, should therefore regularly check and analyse the legal situation in its specific country.

Architecture and Design

Pilot use cases 1 and 2 of the V4Design project focus on architecture and design which belong to the broad spectrum of fine arts. It is common understanding that the freedom of art (Article 13 of the Charter of Fundamental Rights of the European Union) might establish legitimate interest in the processing of personal data. A creative photographer, for instance, might be entitled to take pictures of individuals without their personal consent. It is, however, currently very unlikely that work regarding the two pilot use cases will involve or even depend on the processing of personal data.

Game creation

The same refers to the third pilot use case. PUC 3 focuses on 3D-models of TV series and videos for VR video game creation. Game creation as such belongs to the broader spectrum of fine arts as well. However, it is currently not foreseen by the V4Design project to process any personal data in the context of creating a VR video game.

The consortium will most probably not be able to refer data processing activities back to the freedom of art.

Journalistic storytelling

Work for pilot use case 4, on the other hand, might very well involve the processing of personal data. The focus of PUC 4 is on reliving an event or a specific situation in a 3D or VR environment. The goal is to transform 2D (archive) material into an immersive 3D or VR experience. This original 2D footage as well as additional material that is used to enhance the original footage might include personal information such as images, audio files, names,

dates, metadata, GPS data and other personal data. It will hardly be possible to obtain personal consent from all, if any of the involved data subjects. Consequently, the consortium will need to rely on an alternative legal argument for a lawful processing of this personal data.

This legal argument is the so-called *media privilege*. Its background is the fundamental principle of *freedom of the press* (see Article 11 (2) of the Charter of Fundamental Rights of the European Union). Journalists permanently need to perform research, verify information, analyse and contextualise this information and eventually publish it as news. All these activities involve and actually depend on the processing of personal information and data. Consequently, journalistic work would not be possible if journalists had to observe data protection rules to the same extent as anyone else. The GDPR acknowledges this exceptional situation for journalists and the press in general by an opening clause in Article 85 (1) GDPR:

Member States shall by law reconcile the right to the protection of personal data pursuant to this Regulation with the right to freedom of expression and information, including processing for journalistic purposes.

This means that the consortium needs to refer to the law of the member states in order to assess the extent of media privileges in the context of data processing. As the German international broadcaster Deutsche Welle will be the responsible partner in this use case, the present analysis will focus on German law. Section 12 of the Press Code of North Rhine-Westphalia (Deutsche Welle is based in NRW) as well as Section 9 of the German Interstate Broadcasting Agreement specify the media privilege:

- Journalists are not entitled to use personal data that they have processed in the context of their work for other than journalistic purposes ("data secrecy").
- Journalists who process personal data need to be committing to observe the principle of data secrecy during and beyond their journalistic work.
- Journalists and the press in general are not obliged to observe the provisions of the GDPR apart from
    - the obligation to process personal data in a manner that ensures their integrity and confidentiality (Article 5 (1f) GDPR);
    - the obligation to install a data protection officer whose task is to implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with the GDPR (Art 24 GDPR);
    - the obligation to implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:
        - § the pseudonymisation and encryption of personal data;
        - § the ability to ensure the on-going confidentiality, integrity, availability and resilience of processing systems and services;
        - § the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident;

§ A process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing.

This means that Deutsche Welle is exempt from observing most of the GDPR's provisions *as long as personal data is processed for journalistic purposes*. This assessment can become relevant for V4Design if and as long the stories that will be created by using the V4Design authoring tool are intended for being published in a journalistic context. This particularly refers to the later exploitation and professional usage of the V4Design components. As long as these stories are created purely for testing and evaluation purposes, the legal foundation for the processing of personal data will be legitimate interest (scientific research).

6.1.8 **10 statements regarding data protection within V4Design**

The analysis of European and national data protection regulation has shown that the processing of personal data within V4Design can be realistically based on (1) personal consent, (2) legitimate interest (scientific research) and possibly (3) the media privilege. This leads to the following 10 main statements regarding data processing within the project:

1. *Any handling of data that can theoretically be related to an individual person needs to be considered as processing of personal data which is subject to the provisions of the General Data Protection Regulation.*
2. *Any processing of personal data within V4Design should be based on the data subject's personal consent. This particularly applies when the purpose of data processing is to contact and communicate with specific persons. The consortium will use a standard consent form.*
3. *In case it proves impossible or would involve a disproportionate effort to obtain personal consent, the processing of personal data that is necessary for scientific research within the project can be based on legitimate interest as long as the scientific research interest is not overridden by the interests or fundamental rights and freedoms of the data subject. The latter might particularly refer to personal information about minors and very sensitive data.*
4. *Any personal data should be processed, stored and shared in a manner that safeguards its integrity and ensures that this data does not become accessible to third parties.*
5. *The use of a cloud service is permissible as long as this service complies with the provisions of the GDPR[9].*
6. *Mailing lists need to be used in a manner that one member of the mailing list is not able to access information about the other members.*
7. *Any personal data that is processed for scientific purposes should be anonymised unless this contradicts the scientific research purpose.*

[9] *Comment: Google Drive (at least in its gratis version) is currently not GDPR-compliant.*

8. *Personal data can only be shared with third parties if based on the data subject's explicit personal consent. This also refers to personal data that is processed on the basis of legitimate interest (scientific research).*
9. *Each consortium partner that processes personal data needs to comply with the GDPR compliance rules (see Section 6.1.6.). This includes the involvement of the consortium member's data protection officer.*
10. *The legal and ethical requirements in this deliverable are carefully compiled, but do not constitute any claim with regard to up-to-date material, correctness of the content or completeness. Each consortium partner is legally responsible for all their processing of personal data. This includes the compliance with European as well as national data protection laws.*

6.2 Copyright Law

Unlike the protection of privacy, the protection of copyright has hardly been harmonised on a European level. In fact, one of the guidelines in copyright law is the *Principle of territoriality*. This means that the applicability of national copyright acts is restricted to the respective country's territory. However, the fact that works which are protected by copyright (e.g. texts, films, music, photographs) can be and are used anywhere in the world, has led to a certain alignment of copyright legislation and jurisdiction. This alignment is further supported by a number of international treaties, such as the *Berne Convention for the Protection of Literary and Artistic Works* from 1886. This convention establishes a system of *equal treatment* by requiring that the copyright of works of authors from other parties to the convention (known as members of the Berne Union) are treated at least as well as those of its own nationals. It also requires from its member states certain minimum standards of copyright protection. A number of European directives (starting with the *Computer Programs Directive* in 1991) have contributed to a further alignment of copyright law within the European Union.

6.2.1 General provisions

Without elaborating all aspects of the Berne convention or relevant European directives, it is possible to identify a number of general principles and provisions that apply to most copyright laws in Europe:

- *Protected works*: Copyright protection is granted to authors of *literary and artistic works*. These works include every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression, such as books, pamphlets and other writings; lectures, addresses, sermons and other works of the same nature; dramatic or dramatico-musical works; choreographic works and entertainments in dumb show; musical compositions with or without words; cinematographic works to which are assimilated works expressed by a process analogous to cinematography; works of drawing, painting, architecture, sculpture, engraving and lithography; photographic works to which are assimilated works expressed by a process analogous to photography; works of applied art; illustrations, maps, plans, sketches and three-dimensional works relative to geography, topography, architecture or science (Articles 1 and 2 of the Berne Convention).

- *Individuality*: Most legislation requires an additional qualitative element for copyright protection. This additional element appears in various wordings such as *personal intellectual creation* (German Copyright Act), *intellectual creation with an individual character* (Swiss Copyright Act) or *original intellectual literary, artistic or scientific creation* (Greek Copyright Act). The common ground, however, is that the benchmark for this additional element is rather low. Its sole purpose is to exclude very simple everyday products from copyright protection. In case of doubt, the V4Design consortium should always assume that the work in question is protected by copyright.
- *Duration*: The rights of authors are protected within their lifetime and for seventy years after their death (Article 1 of the Copyright Duration Directive, 93/98/EEC of 29 October 1993). The copyright laws of Belgium (Article 2 of the Law on Copyright and Neighbouring Rights), Germany (Section 64 German Copyright Act), Greece (Article 29 of the Greek Law regarding copyright, related rights and cultural matters), the Netherlands (Article 37 of the Auteurswet), Spain (Article 26 of the Consolidated Text of the Law on Intellectual Property), Sweden (Article 43 of the Act on Copyright in Literary and Artistic Works), Switzerland (Article 29 of the Bundesgesetz über das Urheberrecht und verwandte Schutzrechte) and the UK (Article 12 of the Copyright, Designs and Patents Act) stipulate - with variations - the same copyright duration.
- *Moral and commercial rights*: Copyright legislation in Continental Europe grants copyright owners extensive moral and commercial rights regarding their work. This means, for instance, that their copyright ownership needs to be acknowledged whenever their work is used by other parties (moral right). But it also means that copyright owners have the exclusive right to - commercially or non-commercially - use or exploit their work. This includes the right *to copy*, *to distribute*, *to make available (online)*, *to broadcast* or *to perform* the work, to name just a few. The latter (not the moral rights) can also be licensed to other parties who - in general - are obliged to pay an adequate remuneration for this license. Consortium partners should be aware that the mere storage of this content is a copyright related activity and will, in general, require a valid copyright license.
- *Adaptations and free use*: This aspect might become relevant when video footage or images are transformed into a 3D environment. According to most European legislations (e.g. Section 23 German Copyright Act), adaptations or other transformations of the work may be published or exploited only with the consent of the author of the adapted or transformed work. Only scenarios where the original work fades and is hardly recognisable anymore can be considered as *free use* that does not require any prior consent (Section 24 German Copyright Act). However, the reproduction of a two-dimensional work in a three-dimensional appearance cannot be considered as free use. In fact, the UK Copyright, Designs and Patents Act 1988, for instance, provides in Article 17 that *in relation to an artistic work [unlawful] copying includes the making of a copy in three dimensions of a two-dimensional work*. This finding has no immediate relevance for the datasets that are referred to in this deliverable. However, the consortium needs to be aware that the transformation of footage and images as well as of the motifs they depict into a three-dimensional environment without the copyright owners' consent might by itself establish a copyright infringement.

- *Copyright limitations*: Although copyright legislation grants copyright owners extensive rights, it also acknowledges the role of literary and artistic works for the cultural, scientific and technological development of societies. The creator of a work might be the copyright owner but his or her work also becomes part of the cultural heritage and should - in special circumstances - also be available to other users (even without the copyright owner's permission or any remuneration). One example for these legal limitations is the *right to quote* an extract of a published work to support (e.g. in a scientific context) a position or argument of the person making the quotation. Other examples are the use of works *in news coverage about current affairs* or the usage of works for *private purposes only*. This deliverable is not the place to enlist all existing legal limitations that considerably vary from country to country anyway. However, we will refer to specific legal limitations when assessing the lawfulness of using specific copyrighted works within the V4Design project.

6.2.2 **Texts and Copyright**

Project work within WP2, WP3 and WP5 involves the copying and processing of pre-existing tests that originate from third parties. It is safe to assume that the majority of these texts are protected by copyright. This means that the right to copy as well as to eventually distribute these texts to others (even consortium partners) needs to be licensed from the copyright owner or based on a legal copyright limitation as mentioned in section 6.2.1. One possibility is to license a large number of texts from consortium partners or third parties. In these cases, it will be possible to negotiate a (free) license to use these texts for a certain period of time within the V4Design context. Another possibility would be to access textual content via APIs that grant a simple license to further process this content.

The legal situation becomes more complex if the copying and further processing of texts is based on scraping various heterogeneous sources (e.g. social networks). In these cases, it will not be manageable to contact all authors and to acquire respective licenses. Some of the social networks or other large-scale content providers might grant the right to further use content that was posted on their platforms. However, there is no guarantee that this content was posted by or in agreement with its original author. Social networks cannot effectively grant any license if the original author's consent is lacking

In the light of these legal challenges, a number of legislations have introduced new *copyright limitations* in order to support *scientific research* and *text and data mining*. Section 60c of the German Act on Copyright and Related Rights, for instance, stipulates that isolated articles and small-scale works *may be reproduced for personal scientific research*. This allows for identifying and copying individual shorter texts and articles for research purposes. This includes manual or automated translation or summarisation of these texts.

Section 60d of the German Act on Copyright and Related Rights, on the other hand, supports large-scale text and data mining for non-commercial purposes:

> *In order to enable the automatic analysis of large numbers of works (source material) for scientific research, it is permissible*
>
> *1. to reproduce the source material, including automatically and systematically, in order to create, particularly by means of normalisation, structuring and categorisation, a corpus which can be analysed and*

2. to make the corpus available to the public for a specifically limited circle of persons for their joint scientific research, as well as to individual third persons for the purpose of monitoring the quality of scientific research.

Regarding V4Design, it would be lawful according to German law to create and process large corpora of texts and to share these within the consortium for scientific and non-commercial purposes. However, these corpora cannot be published openly or shared outside the specific research community.

The UK legislation uses the concept of *fair dealing* in order to privilege scientific research. According to Article 29 of the Copyright, Designs and Patents Act 1988, *fair dealing with a work for the purposes of research for a non-commercial purpose does not infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement*. Additionally, Article 29a of the Act provides that *the making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that [...] the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose*.

Other countries, such as Spain, Sweden or Switzerland, have not implemented a similar copyright limitation for scientific research and Big Data analysis yet. This situation is likely to change when the planned Directive on Copyright in the Digital Single Market (2016/0280(COD)) has been passed by the European institutions. The European has formulated its position on September 12th, 2018 and the so called Trilogue-negotiations between Parliament, Commission and European Council are expected to be finalised in 2019. According to the current wording (Article 3 of the Directive), member states shall provide for a copyright limitation for reproductions and extractions of works [...] to which research organisations have lawful access and [that are] made in order to carry out text and data mining for the purposes of scientific research by such organisations. These datasets need to be stored in a secure manner, for example by trusted bodies appointed for this purpose[10].

However, as long as this directive has not entered into force and has not been implemented into national legislation, the use of protected works for scientific research or Big Data analysis remains legally challenging. **It is recommended that each consortium partner that performs this kind of activities, thoroughly examines the legal situation in its own country.**

6.2.3 Images and Copyright

The consortium will create several datasets of *images* (WP2). These images contain architectural drawings, photographs, sketches, paintings, pictures of objects, etc. They will also depict artefacts, buildings, landscapes and will consist of upwards of buildings, architectural features and interior objects. After annotation, a part of this data will be further used for the training and testing of the visual analysis algorithms of WP3 and WP4. As long as this data will become available through the Europeana API, all objects in this dataset are openly licensed. This means that they either fall under the Public Domain or are

[10] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2018-0337+0+DOC+PDF+V0//EN, last accessed on 22/10/2018

licensed under Creative Commons. Consortium partners need to observe eventual special requirements and restrictions regarding the creative commons licenses.

If these images originate from other sources (e.g. Wikipedia, Social Media), the terms and conditions of usage need to be checked on a case by case basis. This refers to the copyright regarding the image itself (e.g. of the photographer or the painter). But it also refers to the image's motif that could be protected by copyright as well. Consortium partners should be aware that *all photographs* (even common snapshots) are protected by copyright. This means that any usage of a photograph that shows a work of modern design or architecture depends on the acquisition of a respective license from the photographer as well as from the designer or architect of the depicted work. In this context, please also refer to the *freedom of panorama* according to section 6.2.6.

6.2.4 **Paintings and Copyright**

The consortium will also create a dataset of *paintings* (WP2). This dataset will contain image files that depict artistic paintings of a certain historical art style. The goal of this dataset is to train the aesthetic extraction algorithms in order to recognise and tag unknown paintings with the art style or styles associated with them, based on this dataset as a train (and test) set. As long as this data will become available through the Europeana API, all objects in this dataset are openly licensed. This means that they either fall under the Public Domain or are licensed under Creative Commons. Consortium partners need to observe eventual special requirements and restrictions regarding the creative commons licenses.

6.2.5 **Videos and Copyright**

The consortium will create several datasets of videos. These videos include

§ full-quality documentaries produced by SLRS, stills of key frames of the same documentaries, and other metadata associated with these documentaries. They are to be used for tasks in WP3 and WP4.

§ full-quality videos stored by AF and metadata associated with these videos (also to be used for tasks in WP3 and WP4).

§ videos coming from DW’s API and video footage.

§ YouTube videos related to the pilot use cases. This dataset will contain the title and the description of the video while account information will be ignored.

When using video content from V4Design consortium partners, other partners need to observe the legal restrictions as laid down in the V4Design Consortium Agreement (esp. Attachment 1: Background included). Video content from other sources (e.g. YouTube) will in nearly all cases be protected by copyright. This means that any usage requires a respective license from the copyright owner.

6.2.6 **Architecture and Copyright**

As already established, the consortium needs to pay special attention to photographs, images and videos of works that are protected by copyright. The depiction of protected works generally requires a respective license by the creator of the depicted work. Most legislation, however, recognise the so-called freedom of panorama as a copyright limitation.

The underlying idea is that buildings or other works (e.g. monuments) that are permanently located in public roads and ways or public open spaces, can be reproduced and distributed without the copyright owner's consent. Whilst the Netherlands and the UK grant this privilege for all premises that are open to the public (see Section 62 of the UK Copyright, Designs and Patents Act 1988), most other legislations are less generous. In Germany, for instance, the privilege only refers to the external view (the facade) and only to footage/images that are taken from the passer-by's perspective. This means that the freedom of panorama does not apply to any footage/images that are taken from helicopters, UAVs and camera cranes or from other positions that are not accessible for the public.

As much as the general idea of the freedom of the panorama is recognised in most legislation, the extent of this privilege differs from country to country. Most legislation allows the private use of footage and images. But the legal situation is far more diverse when it comes to professional use and exploitation. European law allows the member states to individually regulate exceptions or limitations for the use of works, such as works of architecture or sculpture, made to be located permanently in public places (Article 5 lit. h of the Directive 2001/29/EC). Consequently, member states have come to different solutions regarding the freedom of panorama. Germany (Section 59 Copyright Act), Spain (Article 35 section 2 of the Royal Legislative Decree No. 1/1996) allow for professional and profit-driven usage as well. Swedish and Belgian courts, on the other hand, examine on a case-by-case basis whether the usage of footage/images unreasonably affects the copyright owner's legitimate interests. The Swiss Copyright Act provides a variation of the freedom of panorama that is particularly relevant for V4Design. According to Section 27 (2), the privilege does not apply to reproductions in 3D. It is recommended to further analyse this restriction before initiating 3D reproductions within Switzerland. The same refers to the Netherlands where the freedom of panorama does not apply to three-dimensional reproductions. However, the given examples (Memorie van toelichting 28,482 nr 3 (PDF), page 52) such as the reproduction of a building as a souvenir or in a snow globe are not necessarily comparable to the intended use within V4Design where 3D reproduction takes place in a virtual environment.

Very strict is the legislation in Greece. The Law No. 2121/1993 on Copyright, Related Rights and Cultural Matters itself provides in Article 26 that *the occasional reproduction and communication by the mass media of images of architectural works, fine art works, photographs or works of applied art, which are sited permanently in a public place, shall be permissible, without the consent of the author and without payment*. However, Greece has also passed a so-called Antiquity Law that aims for protecting archaeological sites and monuments the country. This means that anyone who intends to take a photograph of ancient Greek sites and re-use these freely has to get a license from the Greek government and pay the set fees (Ministerial Decision ΓΔΑΠΚ/ΔΜΕΕΠ/Γ2/Φ51-52-54/81397/2199). Greek partners in the consortium (CERTH, AUTH) should check whether these restrictions also refer to images from ancient sites that have been created by others and are purely repurposed in the V4Design context.

6.2.7 **Twitter posts and Copyright**

A collection of Twitter posts will be used within WP2. In most cases, the texts in these posts do not show the level of individuality that is needed for copyright protection. However,

copyright protection is possible the more elaborated these texts are (e.g. excerpts from literary texts, poems). Consortium partners also need to be aware that images, videos and other content that are embedded in twitter posts, will most likely be protected by copyright.

6.2.8 **10 statements regarding copyright in V4Design**

The analysis of European and national copyright legislation has shown that the legal situation considerably varies from country to country. Consequently, it will be challenging for the V4Design consortium to comply with the different legal concepts. The following 10 main statements regarding copyright within the V4Design project deliver a rough overview. However, current and future legislation and legal practice need to be frequently checked by all consortium partners on a case-by-case basis.

1. *Copyright legislation is territorial. In principle, each country treats all activities related to copyright according to its own laws and regulations.*
2. *A European Directive on Copyright in a Digital Single Market is in the making and likely to be passed throughout the course of the V4Design project.*
3. *Texts, images, videos and other content items should in principle be considered as protected by copyright as long as the author has not deceased more than 70 years ago.*
4. *All photographs (including snapshots) should be considered as protected by copyright.*
5. *Any processing and use of these content items, including the mere storage in a database require a respective license from the copyright owner or another legal foundation.*
6. *In some legislations (e.g. UK, Germany), works that are protected by copyright can be used - within narrow limits - for scientific research and Big Data analysis without the copyright owner's consent. Each consortium partner should carefully examine the legal situation in its own country.*
7. *The copyright limitation "freedom of panorama" that privileges the reproduction of buildings or other works that are permanently located in public places varies considerably from country to country. Each consortium partner should carefully examine the legal situation in its own country.*
8. *In general, the author has the right to be acknowledged as the author of the work. This means that at least in those cases in which a work that is protected by copyright is singled out and presented or distributed (e.g. in papers, presentations), the work should bear a designation of authorship by use of the symbol "Ó", followed by the author's name.*
9. *The V4Design needs to pay special attention to the broad restrictions regarding the use of footage and images from ancient sites and monuments in Greece.*
10. *The legal and ethical requirements in this deliverable are carefully compiled, but do not constitute any claim with regard to up-to-date material, correctness of the content or completeness. Each consortium partner is legally responsible for*

all their processing of content items that are protected by copyright laws or other legislation on a national or an international level.

6.3 Other intellectual property

The V4Design datasets might theoretically affect some other areas of intellectual property law. The following sections will very briefly discuss the relevance of *design law* and *trademark law*. Actual conflicts with these areas of law need to be examined on a case by case basis.

6.3.1 Design law

As the project bears the word "design" in its acronym (V4Design), it seems obvious to at least briefly examine the implications of design law in the project. Although the faculty of design distinguishes a number of design categories such as industrial (or product) design, communication design or interface design, the area of *design law* mainly focuses on product design. Legally, a product design constitutes the ornamental or aesthetic aspect of a product. More specifically, product design means *the appearance of the whole or a part of a product resulting from the features of, in particular, the lines, contours, colours, shape, texture and/or materials of the product itself and/or its ornamentation* (Article 3 (a) of the Council Regulation (EC) No 6/2002).

Different to copyright protection that does not require any formal act (apart from the creation of the work); design protection is generally based on the designer's application and the subsequent registration of the design. Usually, designs are registered on a national level by national authorities but the EU has also implemented the concept of a *Community Design* to which uniform protection is given with uniform effect throughout the entire territory of the EU. It consequently requires research in national, European and international design registers in order to assess an eventual infringement.

The infringement of a registered design requires an unlawful usage of the design. This implies *the making, offering, putting on the market, importing, exporting, using of a product in which the registered design is incorporated or to which it is applied, or stocking such a product for those purposes* (Article 19 of the Council Regulation (EC) No 6/2002). Consequently, work within V4Design would only violate registered designs if these designs were "used" in the aforementioned sense.

It is indeed very likely that the V4Design consortium will process registered designs and feed them into the authoring tool. However, in most of the cases this will refer to a mere image of the design only which might then be three-dimensionally reproduced in a virtual environment. As long as this process does not lead to a new *product,* it cannot be considered as legally relevant "usage" according to design law.

6.3.2 Trademark law

Similar to the protection of industrial designs, trademarks can also be registered at national level as a national trade mark or at EU-level as a European Union trade mark. According to Article 4 of the Council Regulation (EC) No 40/94, *a Community trademark may consist of any signs capable of being represented graphically, particularly words, including personal names, designs, letters, numerals, the shape of goods or of their packaging, provided that*

such signs are capable of distinguishing the goods or services of one undertaking from those of other undertakings.

Again, it is likely that the footage or other content that will be used and processed within V4Design will depict existing (and protected) trademarks. However, a violation of a trademark requires a usage *as a trademark in the course of trade*. The mere reproduction of a trademark in a virtual environment does not constitute any trademark violation.

6.4 Legal consequences for V4Design datasets

Appendix E delivers an overview of the main legal requirements regarding each dataset that will be created throughout the course of the projects. These requirements are described on a very general level. Please refer to sections 5.1 to 5.3 for further specifications. The consortium should be aware that the statement "no legal restrictions" for some datasets is only valid as long as these datasets do not contain any personal data or any content that is protected by copyright.

7 CONCLUSIONS

This deliverable aims to give extensive insight into the past ten months of work done by the V4Design partners in Work Package 2 of the project. It gives a high-level description of the requirements, the work process, and the outcomes of the initial dataset. This deliverable also works as a reflexive document, incorporating feedback and discussions of the past ten months. It can and should be used as a guideline for the creation of the final dataset in D2.3. Furthermore, the legal and ethical requirements section of this deliverable should inform the policies and work processes of the rest of the project from here on out.

Over the following year, one priority for Work Package 2 lies in creating the technical infrastructure that will allow for automated collection of raw data. Another priority is creating agreements and connections with external 3D cultural heritage object providers.

Integrating the data collection process into the overall V4Design system architecture will require a unified delivery system of data in a unified data format and data model. Currently desk research seems to point towards the SIMMO data model[11] as the most useful for the purposes of the project. Using JSON as the data format for provision of metadata seems to be the easiest and most robust way of sharing metadata. Where applicable, automated scripts will be written that connect to the APIs of data provider partners to pull in data periodically. Where a data provider doesn't have an API, relevant data will be periodically dumped into the V4Design central data storage. Most of the work in the coming year will be geared towards mapping current data provider data models to the SIMMO model, and constructing, testing and integrating the automated scripts into the V4Design system architecture.

Secondly, to be able to provide more 3D objects to architects, video game creators and designers, V4Design partners will seek out active collaboration with external content providers like SketchFab and Scantheworld. Making their 3D objects accessible in the V4Design platform will potentially increase the user engagement with the platform, and promote easy reuse and remixing of those objects.

Lastly, once the technical infrastructure is set up and an agreement is reached with external providers, work will start on forming the final V4Design dataset that will form the foundation of content for the V4Design platform. This final dataset will be delivered together with D2.3.

[11] "A Unified model for Socially interconnected multimedia-enriched Objects", Tsikrika et al., CERTH, 2015, https://github.com/MKLab-ITI/simmo/blob/master/MMMpaper.pdf, last accessed on 22/10/2018

8 REFERENCES

[1] Doe, J., Other, A. N., Another, K. 2011. “A very important paper with wide implications for all activities”, Journal of Very Important Results, vol 2 (35), p. 350-380.

[2] Azevedo, Ana Isabel Rojão Lourenço, and Manuel Filipe Santos. 2008. “KDD,

[3] SEMMA and CRISP-DM: A Parallel Overview.” IADS-DM.

[4] Kurgan, Lukasz A., and Petr Musilek. 2006. “A Survey of Knowledge Discovery and Data Mining Process Models.” The Knowledge Engineering Review 21 (1): 1–24.

[5] Wirth, Rüdiger. 2000. “CRISP-DM: Towards a Standard Process Model for Data Mining.” In Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 32.

[6] Nielsen, Jakob Isak. 2007. Camera Movement in Narrative Cinema: Towards a Taxonomy of Functions. http://www.forskningsdatabasen.dk/en/catalog/2389297223.

[7] Jia, Jiaya, W. Tai-Pang, Yu-Wing Tai, and Chi-Keung Tang. 2004. “Video Repairing: Inference of Foreground and Background under Severe Occlusion.” In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., 1:I–I. https://doi.org/10.1109/CVPR.2004.1315055.

A Appendix A: List of requirements

A.1. Content user requirements

PUC 1: Architectural design, related to existing or historical buildings and their environments

Content user requirements	corresponding HLUR	Description
CUR_001	HLUR_001	videos and images of old and new buildings and landscapes, urban or natural, for 3D extraction
CUR_002	HLUR_002	2D images of old and new buildings and landscapes of which textures and patterns can be extracted
CUR_003	HLUR_006	2D videos/images, textual information, etc. to be used as assets

PUC 2: Architectural design, related to artworks, historic or stylistic elements

Content user requirement	corresponding HLUR	Description
CUR_004	HLUR_007, HLUR_008	videos and images of artworks and culturally sensitive space elements for 3D extraction, aesthetic extraction, and texture and pattern extraction

PUC 3: Design of virtual environments, related to TV series and VR video games

	High LvL User Requirement	Description
CUR_005	HLUR_012, HLUR_013	full episodes of Nico's Weg and corresponding textual exercises
CUR_006	HLUR_019	2D content and textual metadata that can be reused for the design of new, immersive VR environments for language learning purposes

PUC 4: Design of virtual environments, related to actual news for VR (re-) living the date

Content user requirement	High LvL User Requirement	Description
CUR_007	HLUR_020	2D image and video content of environmental objects(e.g. trees, vehicles, landscapes)
CUR_008	HLUR_024	textual metadata for the 2D image and video content that can be attached to the extracted assets
CUR_009	HLUR_025	2D image and video content depicting significant past events that can assist in the design of new, immersive VR environments for reliving a significant past event

A.2. Technical user requirements

VIDEO

content technical requirement	Description
CTR_001	objects filmed from multiple angles for 3D reconstruction
CTR_002	shots containing camera movement with a wide enough baseline for 3D reconstruction
CTR_003	minimal occlusion of objects in shots for 3D reconstruction
CTR_004	relevant metadata describing the action and contents of video
CTR_005	videos delivered in mp4, WAV or RAW format
CTR_006	video timestamps with associated timecodes delivered together with video files
CTR_007	video content contains objects relevant to the PUCs, identified by the keyword list
CTR_008	videos in 720p (HD) quality or higher

IMAGES

Content technical requirement	Description
CTR_008	2D reproductions of canonical paintings of historical artists
CTR_009	images in JPG, PNG, or TIFF format
CTR_010	images of the same object taken from different angles for 3D reconstruction
CTR_011	ground truth tags describing the artistic style or genre of a painting for aesthetic extraction
CTR_012	images showing evenly lit patterns or textures of historical objects or artefacts for texture and pattern extraction

TEXT

content technical requirement	Description
CTR_013	text in the languages covered by the V4Design project: English, French, German, and Greek
CTR_009	text describing or critiquing architectural objects or artefacts
CTR_010	long text descriptions of objects for lexical modelling
CTR_011	shor text descriptions of objects for entity tagging

METADATA

content technical requirement	Description of Action (DoA)
CTR_013	tombstone metadata fields for every DCHO, i.e. - title - description - creator

	- date (of creation)
CTR_014	expanded metadata fields describing the DCHO, i.e. - provenance - licensing and copyright restrictions - geographical location - linked entities - associated entities, e.g. people, places, events
CTR_015	technical metadata fields, i.e. - file format - file size - image dimensions - digitisation gear used
CTR_016	metadata in XML or JSON format

B Appendix B: List of components in the initial dataset

B.1. Iteration 1 phase 1

#	Dataset name header	Date uploaded	PUC #	Partner	Components	Description
001	Hans Scharoun files	22-2-2018	1	EF	63 freely reusable objects, with metadata and image files	query = Hans Scharoun
002	M+ Kowloon files	22-2-2018	2	EF	838 freely reusable objects, with metadata and image files	query = Hong Kong OR Kowloon OR Victoria Harbor OR Victoria Harbour
005	Japanese Art in photos	06-03-2018	2	AF	1200 Japanese photos	approx 1200 photos from Japan. HR versions are available once suitable images are chosen
006	Japanese videos	06-03-2018	2	AF	Films with Japanese art	Copyright permissions are possible after choosing the right images and/or footage
007	Berlin Wall files freely reusable	22-2-2018	4	EF	65 freely reusable objects, with metadata and image files	query = Berlin Wall OR Berliner Mauer OR Antifaschisticher Schutzwall OR Wall of Shame OR Checkpoint Charlie OR Friedrichtstrasse OR Friedrichstraße OR Glienicke Bridge OR Glienicke Brücke OR Oberbaumbrücke OR Dreilinden OR Drewitz
008	general architecture and design video dataset	3/02/2018	1,2,4	AF	Films on chinese design and architecture, interviews, documentaries	CHINA films, Chinese architecture, ornaments, objects and artworks. Interviews with Designers: object design + objects, gallery interior Interviews with

						Designers: furniture Interviews with Designers: interior design, objects and furniture Documentary about street art, featuring buildings
009	SLRS database	27/01/2018	1,2,4	SLRS	4 films from the Jesper Wachtmeister filmography (architecture documentaries).	Microtopia – a documentary about micro dwellings, downsizing and living off the grid, Great Expectations – A Journey through the history of Visionary Architecture, Kochuu – Japanese Architecture / Influence & Origin (better quality will be uploaded), Bruno is Back – a documentary about the Swedish furniture designer and glasshouse architect Bruno Mathsson, Test Site – love, survival, myth, ritual, music, speed and death in the north american deserts
010	Nicos Weg	17/01/2018	3	DW	230 episodes including trailer	Learning German telenovela, level A1, A2 & B1
011	Chateau de Versailles	2/02/2018	4	EF	402 freely reusable results, including media attachments.	Results for the Query 'Versailles', including historical photos from the garden and interior, sketches of decorations and statues, maps of the surrounding area and the palace, blueprints of the grounds, text about the palace, etc.

012	Bruno Mathsson	2/02/2018	1	EF	11 images and 1 sound clip, not necessarily open for reuse or with attached media.	result for the query 'Bruno Mathsson', because SLRS has video sources of Bruno Mathsson's work
013	Eglise de Sorbonne	2/02/2018	4	EF	small curated set of objects	film clips from DW via Euscreen, sketches, pictures, architectural drawings
014	Notre Dame church in Dijon	2/02/2018	4	EF	46 results, images and text, not necessarily freely reusable or with attached media	Notre Dame Church in Dijon query
015	Hans Scharoun (architect)	2/02/2018	1	EF	99 results, 97 images and 2 video clips, of Hans Scharoun, only in limited re-use or no re-use	query for Hans Scharoun, since Artfilms has an interesting documentary on Hans Scharoun with good travelling camera shots of one of his architectural works, Built with Light documentary on AF
016	Jesper Wachtmeister films	7/02/2018	1,2	SLRS	timestamps of possible interesting data, including timestamps, across 3 documentaries	Jesper's list of architectural work, design, landscapes, etc. in his three movies 'Kochuu', 'Great Expectations', and 'Microtopia'
017	Artfilms	17/02/2018	1	AF	Over 600 short clips, 2-5 minutes	screeners for Contemporary Arts Media collection. Contemporary Arts Media's e-commerce Website has all those screeners in curated way.

Table 2: List of components constituting the first phase of Iteration 1 of the initial V4Design dataset

B.2. **Iteration 1 phase 2**

#	Dataset name header	Date uploaded	PUC #	Partner	Components	Description
001	PUC1 Scenario 1 EF images	30/05/2018	1	EF	920 CHOs	heterogeneous dataset from different countries, different providers, and in different languages. Subject of ancient greek architecture with a focus on Delphi.
002	PUC 1 Scenario 2 EF images	30/05/2018	1	EF	5783 CHOs	heterogeneous dataset from different countries, different providers, and in different languages. Subject of German pre-1950s architecture, focus on Berlin architecture.
003	PUC 1 and PUC 4 AF video content		1, 4	AF	23 video objects	9 architecture related videos, 9 mediterranean focused videos, 5 videos on other topics
004	PUC 2 EF images	30/05/2018	2	EF	2693 CHOs	heterogeneous dataset from different countries, different providers, and in different languages. Japanese and Chinese material design.
005	PUC 2 AF images		2	AF	about 600 CHOs	photos from Japan, China, and other images from the films AF distribute.
006	PUC 2 AF video content		2	AF	8 video objects	1 video object on Balinese art and design, 4 videos on Chinese architecture

						and art, 3 videos on Japanese architecture and art.
007	PUC 3 DW content		3	DW	230 video objects, exercises, screenplay, transcripts	230 video objects of Nicos Weg, Web-based exercises for Nicos Weg, 2 screenplays (A1, B1), and four transcipts
008	PUC 4 EF images	30/05/2018	4	EF	7 CHOs	all of the CHOs on the Gendarmenmarkt.
009	PUC 3 Nico's Weg exercises		3	CERTH	440 webpages	Textual information from Deutsche Welle webpages containing Nico's Weg exercises.
010	SLRS videos and timecodes		1, 2, 4	SLRS	3 video objects and associated timecodes	
011	EF aesthetic extraction paintings	30/05/2018	/	EF	approx. 14.000 images and metadata	collection of pre-1950 painting artworks in Europeana, for aesthetic extraction tasks.
012	EF CHO metadata for textual analysis	30/05/2018	/	EF	approx. 23.000 metadata text items	the associated metadata for all provided CHOs, in JSON format, for textual analysis tasks.

Table 3: List of components constituting the second phase of Iteration 1 of the initial V4Design dataset

B.3. **Iteration 2**

#	Dataset name header	Date uploaded	Task #	Partner	Components	Description
001	symmetrical vase dataset	3/08/2018	T4.3	EF	2262 CHOs	heterogeneous dataset from different countries, different providers, and in different languages. Subject of symmetrical small-scale objects like vases, amphoras, bowls.
002	Sum of all Paintings dataset	30/07/2018	T3.5	EF	approx. 37.000 CHOs	Wikidata-extracted dataset from their SPARQL endpoint containing paintings with at least 1 associated genre or style tag.
003	Newspapers dataset	13/08/2018	T3.1	EF	approx. 6.694.985 digitised newspaper pages and metadata	heterogeneous dataset from different countries, different providers, and in different languages. Subject of 592 different newspaper titles from across Europe, collected by the European Library
004	drone footage		T4.3/ T4.4	AF	8 clips appr. 10 mins each	Eight pieces of footage filmed from a drone were added: A ship, a building, cityscape and industrial port objects were filmed from every angle.

Table 4: List of components constituting iteration 2 of the initial V4Design dataset

B.4. Data from external providers

#	Dataset name header	Task #	Partner	Components	Description
001	Wiki webpages scraping	T2.1	CERTH	314 webpages	Textual content and metadata collected from Wikipedia webpages (311 castles plus three pages addressing PUC4).
002	Wiki images scraping	T2.1	CERTH	663 images	Information from images existing in the aforementioned Wikipedia webpages.
003	Twitter posts scraping	T2.1	CERTH	40073 tweets	Twitter posts from 21 user accounts.
004	Flickr dataset	T2.1	CERTH	6209 images	Images returned using 13 queries on the Flickr API.

Table 5: List of components collected from external providers

C Appendix C: PUC keywords

C.1. PUC1 Scenario 1

- Location and Surroundings
  - Greece
  - Delphi
  - Delphic Landscape
  - Ancient Sanctuary
  - Omphalos of Delphi (The navel of the Earth)
  - Pleistos Valley
  - Mount Parnassus
  - Phaedriades
  - Delphi (modern town)
- Neighbouring Buildings
  - Temple of Apollo (Delphi)
  - Temple of Athena Pronaia (image)
  - Stoa of the Athenians
  - Tholos of Delphi
  - Stadium of Delphi (Ancient Greek Stadium)
  - Ancient Gymnasium of Delphi (Gymnasium)
  - Amphitheatre of Delphi (Amphitheatre)
  - World Heritage Sites of Greece (UNESCO)
  - Altar of the Chians
  - Athletic Statues of Delphi
  - Archaeological Museum of Delphi (Architect: A.Tombazis)
- Usage
  - Archaeological Site
  - Sanctuary
  - Theatre (Ancient Greek Drama)
  - Pythia
  - Pythian Games
  - Amphictyonic League
  - Delphi Economic Forum (Modern Greece)
  - Scenography
  - Pavilion
  - Modern Architecture Pavilion Architecture and Design
- Design Features
  - Peripteral Building (Peripteros)
  - Doric Order , Ionic Order , Corinthian Order
  - Ornaments
  - Landscape Design / Landscape Architecture
  - Capital (Architecture)
  - Columns
  - Excavation (Archaeology)
- Ancient Greek Statues
  - Polygonal Wall
  - Marble
- References
  - Greek Mythology
  - Ancient Greece
  - Trojan War
  - Ancient Theatre (Epidaurus)
  - Dionysus
  - Apollo

C.2. PUC1 Scenario 2

- Location and Surroundings
  - Germany
  - Berlin, West Berlin

    - Kulturforum
    - Staatliche Museen zu Berlin (Berlin State Museums)
    - Potsdamer Strasse / Potsdamer Platz
    - Herbert-von-Karajan-Straße
    - Neue Nationalgalerie (New National Gallery) / 1968 /
    - Ludwig Mies van der Rohe
    - Matthäuskirche (St. Matthew Church) / 1845 /
    - Friedrich August Stüler
    - Berliner Philharmonie / 1963 /
    - Hans Scharoun
    - Staatsbibliothek zu Berlin (Berlin State Library)
    - Hans Scharoun
    - Gemäldegalerie (Berliner Gemäldegalerie)
    - Archer Sculpture / Henry Moore
- Usage
    - Museum / Gallery / Art Museum
    - Collection
    - 20th century Art
    - Exhibition Design
- Design Features
    - Brick
    - Glass Brick
    - Pitched Roofs
    - Landscape Design / Landscape Architecture
- References
    - Tate Modern Extension
    - De Young Museum
    - Alte National Galerie Berlin

C.3. **PUC2**

- Location
    - East Asia
    - Japan
    - Nippon
    - Tokyo
- Related Objects
    - Japanese calligraphy
    - Woodblock Printing
    - Suzuri-bako
    - Sword crafting
    - National treasures of Japan
    - Japanese Craft
    - Traditional Glass production (Japan)
    - Paper making (Japan)
    - Inkstone Carving
    - Japanese Dolls
    - Japanese utensils
    - Japanese ceramics
    - Textiles
    - Chinese Calligraphy
- Usage
    - Exhibition Design
    - Furniture DesignLighting (Lighting Design)
    - Signage (Museum signage)
    - Graphic Design
    - Tapestry
    - Visual identity
    - Museum Products
- Design Features
    - Ornaments
    - Porcelain
    - Bamboo
    - Patterns
    - Kagome Pattern (Japan)

    - Japanese woodworks
    - Japanese bamboo works
- References
    - WestKowloon (Cultural District)
    - M+ Museum
    - Japanese Exhibition House (MoMA)
    - Japanese Constellation (MoMA)

C.4. **PUC3**

For PUC3 the only data collected for the initial dataset were objects related to DW's Nico's Weg telenovela. Because the subject and scope of the data to be collected was very clearly outlined, no keywords were needed for this PUC.

C.5. **PUC4**

For PUC4 the goal was to create a VR application for reliving the date, which would be realised as a collaboration between NURO and DW. The subject of this VR application was first decided to be the Berlin wall, but after discussion within the consortium it was deemed safer to take the Berlin Gendarmenmarkt as a casus. The only content guidelines for data collected for this use case was content connected to the Berlin Gendarmenmarkt, the keywords consisted of Named Entities of the buildings that constitute that plaza and the famous artists that are depicted at and worked on the plaza.

- Französischer Dom
- Deutscher Dom
- Konzerthaus Berlin
- Friedrich Schiller statue
- Johann Arnold Nering
- Georg Christian Unger

D Appendix D: Overview of the legal requirements for each dataset

Dataset reference & name	Dataset description	Main legal requirements
Dataset_WP1_1_ ContactDetails	The database contains contact information for the project partners and the advisory board members.	Data can only be processed on the basis of personal consent. Data needs to be secured so that it is not accessible for third parties. Data needs to be erased 3 months after the end of the project at the latest.
Dataset_WP2_1_ EF_Images	This dataset contains image files that are on the topic of architecture and design (architectural drawings, photographs, sketches, paintings, pictures of objects, etc).	A lot of content in this dataset will be protected by copyright. Consent of the copyright owner is required, e.g. by Europeana. Scientific research might establish a copyright limitation in some countries.
Dataset_WP2_2_ EF_Paintings	This dataset contains image files that depict artistic paintings of a certain historical art style.	Content in this dataset will be protected by copyright, as long as the author has not deceased more than 70 years ago. Consent of the copyright owner is required, e.g. by Europeana. Scientific research might establish a copyright limitation in some countries.
Dataset_WP2_3_ SLRS_Videos	This dataset includes full-quality documentaries produced by SLRS, as well as stills of keyframes of the same documentaries.	Videos in this dataset will be protected by copyright. Consent of the copyright owner of the documentaries (here SLRS) is required (see Consortium Agreement). Motifs that are shown in the documentaries might be protected by copyright as well. Consent of the copyright owners of protected works that are shown in the documentaries is required if no copyright limitation is applicable. Scientific research might establish a copyright limitation in some countries.
Dataset_WP2_4_ AF_Images	This dataset contains image files that are on the topic of architecture and design. These images contain artefacts,	Most of the images in this dataset will be protected by copyright. Consent of the copyright owner is required.

	buildings, landscapes, people, and will consist of upwards of buildings, architectural features and interior objects.	Scientific research might establish a copyright limitation in some countries. Many image motifs will also be protected by copyright. Freedom of panorama as a copyright limitation needs to be checked on the basis of the applicable national legislation. The consortium needs to observe the strict regulations in Greece. Images of people that randomly appear in the content can be processed on the basis of scientific research as legitimate interest.
Dataset_WP2_5_ AF_Videos	This dataset includes full-quality videos stored by AF.	Videos in this dataset will be protected by copyright. Consent of the copyright owner of the documentaries (here AF) is required (see Consortium Agreement). Motifs that are shown in the documentaries might be protected by copyright as well. Consent of the copyright owners of protected works that are shown in the documentaries is required if no copyright limitation is applicable. Scientific research might establish a copyright limitation in some countries.
Dataset_WP2_6_ Wiki_Text	This dataset contains textual information coming from 314 Wikipedia Web pages. The text is the plain text of the Wikipedia Web pages and the info box.	Content in this dataset will be protected by copyright. The terms of the Wikipedia Creative Commons License need to be observed.
Dataset_WP2_7_ Wiki_Images	This dataset contains 670 image files extracted from the aforementioned 314 Wikipedia Web pages. These are mainly images of castles as well as landmarks.	Images in this dataset will be protected by copyright. The terms of the Wikipedia Creative Commons License need to be observed. Depicted castles and landmarks might be protected by copyright, as long as the creator/architect has not deceased more than 70 years ago. Freedom of panorama as a copyright limitation needs to

		be checked on the basis of the applicable national legislation. The consortium needs to observe the strict regulations in Greece.
Dataset_WP2_8_ Twitter_Posts	This dataset contains 40073 Twitter posts published by 21 Twitter accounts.	No legal restrictions as long as the Twitter posts do not show the level of individuality that is needed for copyright protection. no personal data is processed.
Dataset_WP2_9_ DW_Exercises	The dataset contains textual content from Nico's exercises for learning German.	The textual content is protected by copyright. Consent of the copyright owner is necessary (here DW) irs required, (see Consortium Agreement).
Dataset_WP2_10_ DW_Videos	This dataset will contain videos coming from DW's API and video footage.	Videos in this dataset will be protected by copyright. Consent of the copyright owner of the documentaries (here DW) is required (see Consortium Agreement). Motifs that are shown in the documentaries might be protected by copyright as well. Consent of the copyright owners of protected works that are shown in the documentaries is required if no copyright limitation is applicable. Scientific research might establish a copyright limitation in some countries.
Dataset_WP2_11_ YouTube_Videos	This dataset contains YouTube videos related to the pilot use cases. The dataset contains the title and the description of the video while account information will be ignored.	Videos in this dataset will be protected by copyright. Consent of the copyright owner of the videos is required. Motifs that are shown in the videos might be protected by copyright as well. Consent of the copyright owners of protected works that are shown in the videos is required if no copyright limitation is applicable. Scientific research might establish a copyright limitation in some countries.
Dataset_WP2_12_ Flickr _Images	This dataset contains Flickr images related with the pilot use cases.	Images in this dataset will be protected by copyright. Consent of the copyright owner

		of the images is required. required if no copyright limitation is applicable Scientific research might establish a copyright limitation in some countries.
Dataset_WP3_1_ Summarisation	The dataset will consist of texts in the different V4Design languages with their associated summaries, and/or annotations that indicate which parts are relevant to the end users.	Texts in this dataset as well as manually created summaries will be protected by copyright. Consent of the copyright owners of the texts/summaries is required. Scientific research might establish a copyright limitation in some countries.
Dataset_WP3_2_ ConceptExtraction	The dataset will consist of large collections of texts in the different V4Design languages for the different V4Design domains, according to which they will be clustered.	Texts in this dataset will be protected by copyright. Consent of the copyright owner of the texts is required. Scientific research might establish a copyright limitation in some countries.
Dataset_WP3_3_ Aesthetics	The dataset will serve as an annotated benchmark which will be used to train the aesthetics extraction models. It will contain images with the appropriate aesthetics tags. The dataset will be annotated based on their painting style, creator and architecture type.	Most of the images in this dataset will be protected by copyright. Consent of the copyright owner is required. Many image motifs will also be protected by copyright. Consent of the copyright owner is required. Scientific research might establish a copyright limitation in some countries.
Dataset_WP3_4_ TextureProposals	The aesthetics (stored in the Dataset_WP3_3_Aesthetics) will be combined with appropriate aesthetics models to produce novel texture proposals which will be kept in the current dataset.	No legal restrictions as long as the aesthetics models are not protected by copyright or any other law.
Dataset_WP4_1_ Localisation	The dataset will contain videos acquired from WP2. These videos will be further enhanced here with more meaningful metadata, such as Spatio-Temporal Video Localisation tags and 3D models of the buildings and architecture structures that	3D reproduction of protected works is legally restricted in the Netherlands and in Switzerland.

	might exist in each video.	
Dataset_WP4_2_ PhotogrammetryInput	Contains data used as input for the photogrammetry pipeline. Most likely this would consist of extracted frames, reference to source and timestamps.	No legal restrictions as long as no personal data is processed.
Dataset_WP4_3_ PhotogrammetricReconstruction	Contains photogrammetric reconstruction objects. These objects contain all relevant data from the photogrammetric reconstruction pipeline (calibration, matching, point clouds, meshes).	No legal restrictions.
Dataset_WP4_4_ AlternativeReconstuction	Repurpose of input material where no standard multiview reconstruction was used: texture extraction, rectification, etc.	No legal restrictions.
Dataset_WP4_5_ EnrichedModels	Enrichment from Dataset_WP4_2 & dataset_WP4_3 where possible. Models may be further enhanced or segmented ('polished') based on the input from the visual understanding tool and mesh analysis.	No legal restrictions.
Dataset_WP5_1_ AssetAnnotationsDataset	The dataset will contain the semantic representation of the annotations that are generated by the various V4Design modules.	No legal restrictions.
Dataset_WP5_2_ LanguageGenerationDataset	The dataset will consist of texts in the different V4Design languages with superimposed semantic (predicate-argument) annotations aligned with syntactic annotations.	Texts in this dataset will be protected by copyright. Consent of the copyright owner of the texts is advisable. Scientific research might establish a copyright limitation in some countries.
Dataset_WP5_3_ LinkedDataModels	Further enrichment of Dataset_WP4_5_EnrichedModels. To contain BIM like objects.	No legal restrictions.
Dataset_WP6_1_Architecture IntegrationSpecification	Functional and non-functional requirements, hardware requirements, component descriptions (inputs &	No legal restrictions.

	outputs), component dependencies, API descriptions, information flow diagram, internal and external interfaces and testing procedures.	
Dataset_WP7_1_ UserInteraction	The data that is generated by the users of the V4Design platform like users' personal information, detailed log of user actions (login, logout, account creation, visits on specific parts of the tool), information on user devices. By means of these data the system will generate statistics e.g. the 3D model most viewed/used, and other metrics to measure performance and other concerns.	Any processing of personal data in this context needs to be based on the individual user's informed consent. The user's consent can be given online when registering to the V4Design authoring tool (opt-in). The user's consent is subject to withdrawal at any time. Personal data need to be stored in a secure environment that is not accessible to third parties. Personal data needs to be erased as soon as it is not necessary anymore to pursue the original purpose.
Dataset_WP8_1_ Dissemination	The database contains contact information of the dissemination group of the project. This includes members of the network of interest and confirmed members of the user group. It might also include participants of events, recipients of the newsletter etc.	Any processing of personal data in this context needs to be based on the individual user's informed consent. The user's consent is subject to withdrawal at any time. Personal data need to be stored in a secure environment that is not accessible to third parties. Personal data needs to be erased as soon as it is not necessary anymore to pursue the original purpose.
Dataset_WP8_2_ CommunicationsMonitoring	The database is expected to contain information on key communications indicators, including the number of site visits, code downloads etc.; the number of participants to events; (with respect to the Social Media) the followers and engagement; (with respect to the Publication) the number of publications in technical, scientific and academic	Any processing of personal data in this context needs to be based on the individual user's informed consent. The user's consent is subject to withdrawal at any time. Personal data need to be stored in a secure environment that is not accessible to third parties. Personal data needs to be erased as soon as it is not necessary anymore to pursue the original purpose.

	conferences and journals; (with respect to the User Group) the number of users and the diversification The data sources that will be used to retrieve part of the data include: Twitter, Facebook and LinkedIn Analytics. In addition, Matomo has been established to analyse the traffic of the V4Design Website.	

E Appendix E: Feedback form template for evaluation of each iteration

Name					
Affiliation					
Feedback on Datasets					
Please rank from 1 to 5 according to the following scale:					
Scale:	totally unsuitable	1			
		2			
	to be improved	3			
		4			
	perfectly suitable	5			

			Aesthetics extraction & Texture Proposals (AE&TP)		
Visual/Textual dataset			Format	Quality	Comments
		PUC1-Sc1		10/10 (5)	
		PUC1-Sc2		10/10 (5)	
		PUC2			
	Images	PUC4		7/10 (3)	example comment
EF	Paintings			7/10 (3)	
DW					
		Great_Expectations_Solaris	NA	10/10 (5)	
SLRS	Videos	Kochuu_Solaris	NA	10/10 (5)	

		Microtopia_Solaris	NA	10/10 (5)	
	Images	PUC1-Sc2 (Berlin)		10/10 (5)	
		PUC2 (Japan, China, Other images)		7/10 (3)	
	Videos	Architecture related films		7/10 (3)	example comment
		Balinese art film		10/10 (5)	
		China related films		10/10 (5)	
		Japan film		10/10 (5)	
		Mediterranean topics films		10/10 (5)	
AF		Other footage		10/10 (5)	
	Wikipedia images				
	Wikipedia webpages				
Scraping	Twitter posts				

			Spatio-Temporal Building and Object Localization (STBOL)		
Visual/Textual dataset			**Format**	**Quality**	**Comments**
		PUC1-Sc1			
		PUC1-Sc2			
		PUC2			
EF	**Images**	**PUC4**		**7/10 (3)**	**Some images are .jpeg and others .png. It will be better if all of them have the same extension**

	Paintings				
DW					
SfP	Videos	Great_Expectations_Solaris	NA		
		Kochuu_Solaris	NA		
		Microtopia_Solaris	NA		
AF	Images	PUC1-Sc2 (Berlin)			
		PUC2 (Japan, China, Other images)		7/10 (3)	example comment
	Videos	Architecture related films		7/10 (3)	
		Balinese art film		7/10 (3)	
		China related films		10/10 (5)	
		Japan film		10/10 (5)	
		Mediterranean topics films		10/10 (5)	
		Other footage		10/10 (5)	
Scraping	Wikipedia images				
	Wikipedia webpages				
	Twitter posts				

F Appendix F: V4Design standard consent form template

The following standard consent form needs to be used in all circumstances where individuals are asked for their contact details in order to receive newsletters or to be part of the V4Design user group. In case the subscription takes place by using a contact form on the V4Design Website, this consent form needs to appear directly before the contact form. In case of email or other communication, the respective person needs to explicitly consent to the terms as described hereunder.

V4Design standard consent form

Basic principles

We collect, process and use your personal data in compliance with the EU General Data Protection Regulation (GDPR). Personal data refers to all individual pieces of information about personal or factual details of an identified or identifiable natural person. This includes, for example, your name, your address, your e-mail address and your telephone number.

Contact forms, newsletter

You can use the contact forms provided on our Website to contact us directly or to request current information from us. We collect, process and use the information you provide us with in a contact form exclusively to handle your request.

If you have registered for one of our newsletters, we collect, process and use the information you enter exclusively to send out the relevant newsletter. You can unsubscribe from the relevant newsletter at any time.

User group

If you have signed up to the V4Design user group, we collect, process and use the information you provide us with exclusively in order to inform you about the project and its progress on a regular basis, to invite you to V4Design events and to invite you to participate in the V4Design prototype evaluation.

Term of storage

Your data will be stored for no longer than it is necessary for the original purpose(s). It will be erased three months after the end of the V4Design project the latest.

Access and withdrawal of consent

You have the right to free access to any of your personal data that is stored. You can withdraw your consent to collect, process and use your personal data at any time, with effect for the future by sending an email to *XXX[at]v4design[dot]eu*.

We are happy to answer any further questions on data protection and processing of your personal data. You can contact us using the e-mail address: *XXX[at]v4design[dot]eu*.