Grant Agreement No.: 761488

D4.1: Initially available datasets and usage guidelines

Work package	WP 4
Task	T4.2
Due date	28/02/2018
Submission date	28/02/2018
Deliverable lead	VRT
Version	1.0
Authors	Matthias De Vriendt (VRT), Joris Mattheijssens (VRT), Olga Kisselmann (DW), Christos Danezis (DIAS)
Reviewers	Matthias Strobbe (IMEC), Chris Develder (IMEC)
Keywords	CPN, Personalisation, Dataset, Data, Metadata

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	30/11/2017	Draft version	Janssens Bart, Matton Mike (VRT)
V0.2	21/02/2018	Initial version	Matthias De Vriendt (VRT), Joris Mattheijssens (VRT), Olga Kisselmann (DW), Christos Danezis (DIAS)
V1.0	28/02/2018	Final version	Matthias De Vriendt (VRT), Joris Mattheijssens (VRT), Olga Kisselmann (DW), Christos Danezis (DIAS), Charlotte Knapen (VRT)

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

DISCLAIMER

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

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

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

EXECUTIVE SUMMARY

This deliverable is produced in the context of Task 4.1, which aims to organize a meaningful and productive pilot. It collects, integrates and manages the different data sources.

D4.1 will allow us to use machine learning algorithms and tools while implementing the planned personalisation pilots. These pilots serve as self-contained proof-of-concept field trials. Promising pilots may then be presented to the publishing partners as a basis for efforts related to the ongoing push for personalised content and experiences prevalent in the publishing industry today.

The following data will be made available by the appropriate partners (DW, VRT and DIAS) at the very beginning of the project:

- Data structures of articles and coverage of article metadata
- Summary process diagrams on data collection and relevant management processes
- Content consumer profiles and their relevant contextual data

However, to guarantee a smooth project process, data will have to be collected by each broadcasting or publishing partner. We foresee a large data collection exercise for the following data: content (e.g. news articles), user context (e.g. location, device, time of day) and user behaviour.

The broadcasting and publishing partners in the consortium have varying levels of data collection capabilities. Partners will be improving their data collecting and management competences during the lifetime of the project. Thus, we hope to realize a seamless transfer of data from the live production environments of broadcasters and publishers to the CPN platform.

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

TABLE OF CONTENTS

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

ABBREVIATIONS

IP	Internet Protocol
TCP	Transmission Control Protocol
HTTP	HyperText Transfer Protocol
CSV	Comma Separated Values
REST	Representational State Transfer
JSON	Javascript Object Notation
ETL	Extraction, Transformation and Load
API	Application Programming Interface

1. INTRODUCTION

As suggested by the title, we position D4.1 as the description of the initially available datasets and usage guidelines. This deliverable organizes a meaningful and productive pilot. It furthermore aims to define the goals as well as the scope and content of the pilot. A key element is to allow for a reality check of assumed demand from users and calibration of the platform to meet expectations.

This deliverable starts with a definition of the different types of available data. We make a distinction between raw data (facts) and inferred knowledge that comes from this data. A comprehensive listing of each partner's available datasets is be presented, along with a detailed explanation for each field contained in the dataset. In other words, we tend to answer the following questions:

- What does the field represent?
- How is it determined or computed?
- Is it a fact or derived knowledge?

Furthermore, we show a pedigree per dataset, which details how the dataset was collected and compiled. Lastly, we present a way to manage and structure changes to the provided datasets, as more detailed or additional data is collected during the project. Collecting data is an ongoing effort that will have to facilitate the needs of machine learning systems as they evolve and mature. For this purpose, we consider possible ways of versioning available data.

To conclude this deliverable, we deep-dive into some of the most common pitfalls when working on big data problems, as we might encounter these and present possible solutions to them.

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

2. TYPES OF DATA

In order to make each partner's data collection exercise as fruitful as possible, we'll first make a distinction between facts (which we'll call raw data) and derived data such as aggregates, inferred data or assumptions.

Raw data or facts describe events that we know have occurred with 100% certainty. Facts include examples such as:

- A user views an article on the VRT NWS website. We measure his browser tab is opened for 33 seconds.
- A user streams media content from the VRT NU platform. We detected his browser signature is `Mozilla/5.0 (Linux; <Android Version>; <Build Tag etc.>) AppleWebKit/<WebKit Rev>`

Derived data includes aggregated data, or data inferred from the factual data. Building on the above, the derived data can include the following examples:

- A user views an article for 33 seconds, and has scrolled regularly. Therefore we infer the user has spent over 30 seconds being engaged with the content.
- From a user's browser signature, we infer he or she is most likely using an android mobile device.
- Aggregating a specific user's views over a long timespan, we compute his or her article category affinities.

3. COMMON PITFALLS

Having made the distinction between raw and derived data, we can now focus our efforts on the collection of meaningful and accurate raw data. The most important problem we face during this phase of a machine learning project is selecting the appropriate data features that need to be collected to realize a certain goal. In the CPN project, this goal is to facilitate the development and implementation of algorithms and components that personalise the content and experience of content consumers in a media context.

Having clearly described this eventual goal we would like to accomplish, e.g. the ability to accurately estimate the probability a user will enjoy a given article. Or maybe we would like to estimate the optimal time of day for an interaction between a given user and a notification from the VRT NWS android application.

When we have these goals defined, we don't know which data points would make a good set of variables from which we can make an accurate estimate. At this point we can only make an educated guess regarding which data features would help to improve the accuracy of content recommendations made by a machine learning model as we don't have a trained model yet which can be evaluated.

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

Imagine the following scenario: We have a dataset containing the viewing behaviours on a news website per anonymous user. This dataset may contain features such as articles viewed, category of the article, the view duration, the time of day, etc. In order to make better article recommendations we might decide to collect demographic information about our users such as age, sex and maybe even Facebook likes. Due to the regulations enforced by GDPR, this venture might consume a tremendous amount of time and effort. However, even though we may collect many more additional features and data points, this doesn't always mean that this will result in a proportional increase the accuracy and overall quality of recommendations made by a model.

In order to avoid this common pitfall, and use our valuable time as best as possible, we need a way of telling on beforehand if certain additional features will significantly improve the performance of a model. One way of doing this is by performing a ceiling analysis[1]. This can allow us to determine the influence of each individual pipeline component on the overall performance of the system. Then we can focus our effort on the component with the highest potential for increased system accuracy. In the context of a machine learning model used for making content recommendations, this would mean a higher rate of content consumers interacting with, or responding positively, to recommendations made.

[1] Ceiling analysis of pedestrian recognition pipeline for an autonomous car application (http://ieeexplore.ieee.org/document/6521941)

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

4. COLLECTING RAW DATA

4.1 DIAS: DATA COLLECTION PROJECT

In order to achieve content personalization DIAS will collect and make available raw data from the users of the digital news portal **sigmalive.com**. As the organization is currently exploring ways to make data available, it has created a project which consists of the following parts:

1. Collection of Article Information Raw Data

These data describe an article and have not been processed. The collection contains information regarding the title, the path, the published date, and other which are closely related with the identification of a specific article and the elements that describes it. These data will be extracted directly from the CMS.

2. Collection of User Viewing Behavior

In order to collect the data that describes how the user consumes the content, a custom solution software will be implemented. This solution will be able to generate sessions and user IDs for a period of time.

3. Azure Infrastructure Setup

Currently, we are planning to use the Azure infrastructure to collect the massive amount of generated data. However, we are working on other solutions to implement big data collection. By the next reporting period we will be ready to discuss and explain this further.

4.2 DATA COLLECTION METHODS AT VRT

Our goal is to properly collect and process large amounts of data. Therefore we need two key components:

- Infrastructure to ingest or accept new data points to the available set of data;
- Infrastructure to process and clean large amounts of raw data.

Currently, we have set up a big data streaming platform based on Apache Kafka. We introduce new data to the system by using a combination of receiving HTTP POST requests sent to our endpoint by VRT NU, and sending HTTP GET requests to a third party API endpoint providing live traffic data on the VRT NWS website.

The raw data is then normalised and refined using the Kafka Streams API, after which the refined data is then stored on Amazon S3 storage in a CSV format.

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

Since we have open-sourced the components used in our setup[2], these can be used to set up a similar infrastructure without expending a large amount of the required R&D effort.

4.3 COLLECTED DATASETS

Because our goal is to personalize content of an online news outlet for its users in mind, we'll briefly list interesting data points that have been collected by VRT R&D. Each item is accompanied by an example of what its schema looks like.

DIAS and Deutsche Welle have also listed some of the datasets that can be provided to the consortium partners. The possibility to supply additional data sets needed to successfully implement various personalisation algorithms and components, will be reviewed with business executives.

4.2.1 VRT: Article information

The data that VRT is currently collecting are as follows:

Article Metadata	
Id	Unique identifier for the article. Used for joining the article onto other datasets
Title	Title given to the article
Path	url slug for the path pointing to this article
ImageUrls	Urls to images contained in the article, if any
Introduction	Paragraph of text shown on smaller views of the article, for example on the front page.
Subtitle	Subtitle given to the article
publishedDate	Date of publication
Tags	Tags given in the CMS system, category etc.
Types	Indicates if an article contains video, is part of a longer story etc.
Links	URLs pointing to this article

[2] see https://github.com/dataprism for the components that we have published as open source

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

Article Metadata Example	
Id	1519715157142
Title	"Vieze vakantiehuisjes in Nederland: stof en beestjes, maar ook vuiligheid die je niet met het blote oog ziet"
Path	"/content/vrtnieuws/nl/2018/02/27/nederland---bungalows-in-vakantieparken-zijn-niet-proper-"
ImageUrls	"https://images.vrt.be/orig/2018/02/27/c9bc4ac4-1b94-11e8-abcc-02b7b76bf47f.jpg"
Introduction	"Veel bungalows op vakantieparken in Nederland worden niet goed schoongemaakt. Stof, kalk, beestjes, schimmel... Maar ook viezigheid die niet met het blote oog te zien is. Een en ander blijkt uit een onderzoek van de Nederlandse Consumentenbond waarover de NOS schrijft."
Subtitle	Veel bungalows op vakantieparken in Nederland worden niet goed schoongemaakt, zo blijkt uit een onderzoek van De Consumentenbond.
PublishedDate	2018-02-27T08:11:15+0000
Tags	"categoryTags":[{ "tagId":"functional:vrtnieuws/categories/buitenland", "displayName": "Buitenland" }]
Types	"isvideo"
Links	[{ "rel": "self", "url": "https://www.vrt.be/vrtnws/nl/2018/02/27/nederland---bungalows-in-vakantieparken-zijn-niet-proper-.app/" }, ...]

A sample is shown in the Annex to this document.

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

4.2.2 Deutsche Welle: Article information

As a content partner for CPN Deutsche Welle will provide an extensive amount of news articles and media items for the CPN platform as shown in **List of available DW content in all DW languages** in the Annex, as well as metadata for content item identification. The metadata will be provided in Json format and made accessible via DW's REST API. The API closely mirrors the content on the website[3] and gives access to articles, audio, video and image galleries.

Article Metadata	
Type	An indication if the article contains video, or other type of data etc
Id	Unique identifier for the article. Used for joining the article onto other datasets
Language ID	Language Number of Item
Tracing info	Position on the webpage
Name	Title given to the item
Teaser	Paragraph of text shown on smaller views of the article, for example on the front page
Category Name	Category as in CMS
Text	Text of the Item
Permalink	URL to the Item
Display Date	Date of publication

[3] www.dw.com

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

Author(s)	Tags given in the CMS system, category etc.
Keywords	Keywords assigned to item in CMS
Reference Group	Similar/associated items
Image Used	Urls to images contained in the article

Article Metadata example	
Type	"Article"
Id	42516274
Language ID	1 (German)

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

Tracing info	{
	"level2": "1",
	"page": "<prefix>::THEMEN::Wirtschaft::Bitcoin: Meinungen, Mythen, Missverständnisse",
	"customCriteria": {
	"x8": "",
	"x9": "20180212",
	"x10": "<prefix>::THEMEN::Wirtschaft",
	"x1": "1",
	"x2": "1",
	"X14": "",
	"x3": "42516274",
	"x4": "1503",
	"x5": "Bitcoin: Meinungen, Mythen, Missverständnisse",
	"X15": "",
	"x6": "1",
	"X18": "",
	"x7": ""}}

Name	"Bitcoin: Meinungen, Mythen, Missverständnisse",
Teaser	"Am Bitcoin scheiden sich die Geister: Stirbt er oder dominiert er bald sogar die gesamte Finanzwelt? Beide Szenarien sind unwahrscheinlich, sagt der Analyst Jochen Möbert im DW-Interview."
Category Name	"Kryptowährungen"

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

Text	"Deutsche Welle: Sie haben versucht, sich den unterschiedlichen Aspekten der Kryptowährung Bitcoin anzunähern. Zu welchen Ergebnissen sind Sie gekommen? Jochen Möbert: Die Kryptowährungen basieren auf der Blockchain-Technologie. Da sie global und dezentral organisiert ist und auf dem Internet aufsetzt, unterscheidet sie sich stark von vielen anderen Entwicklungen in der Technikgeschichte. Ganz klar hat sie auch das Potenzial, das Bankgeschäft und die Finanzindustrie zu revolutionieren. Trotz der Euphorie, die manchmal in diese Technologie hineininterpretiert wird, steckt sie noch in den Kinderschuhen. Es dürfte noch ein weiter Weg sein, bis wirklich marktreife Produkte und Lösungen entwickelt sind, die das traditionelle Bankgeschäft in neue Bahnen lenken. Gleich am Anfang stellen Sie fest, wie sehr der Bitcoin polarisiert:"
Permalink	"http://p.dw.com/p/2sOQM",
Display Date	"2018-02-12T21:45:00.000Z"
Author(s)	[{"name": "Klaus Ulrich","addendum": ""}],

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

Keywords	{
	"name": "Schlagwörter",
	"type": "Keywords",
	"items": [{
	"type": "SearchRef",
	"name": "Bitcoin",
	"url": "https://api.dw.com/api/search/global?terms=Bitcoin&languageId=1"
	},
	{
	"type": "SearchRef",
	"name": "Kryptowährung",
	"url": "https://api.dw.com/api/search/global?terms=Kryptow%C3%A4hrung&languageId=1"
	}, {
	"type": "SearchRef",
	"name": "Kryptogeld",
	"url": "https://api.dw.com/api/search/global?terms=Kryptogeld&languageId=1"
	}, {
	"type": "SearchRef",
	"name": "Blockchain",
	"url": "https://api.dw.com/api/search/global?terms=Blockchain&languageId=1"
	},}],

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

Reference Group	"referenceGroups": [{
	"name": "Die Redaktion empfiehlt",
	"type": "InternalContent",
	"items": [{
	"id": 42466006,
	"type": "ArticleRef",
	"name": "Drastische Warnung vor Bitcoin - Absturz hält an",
	"url": "https://api.dw.com/api/detail/article/42466006"
	}, {
	"id": 42189051,
	"type": "ArticleRef",
	"name": "Bitcoin-Panik: Crash beim Kryptogeld?",
	"url": "https://api.dw.com/api/detail/article/42189051"
	}, {
	"id": 42107289,
	"type": "ArticleRef",
	"name": "Hindernisse für Bitcoin häufen sich",
	"url": "https://api.dw.com/api/detail/article/42107289"
	}, {
	"id": 41766115,
	"type": "ArticleRef",
	"name": "Die schmutzige Seite des Bitcoin-Booms",
	"url": "https://api.dw.com/api/detail/article/41766115"
	}

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

Image Used	"mainContent": {
	"id": 42504817,
	"type": "Image",
	"name": "Symbolbild Kryptowährung Bitcoin mit Würfeln (picture-alliance/Klaus Ohlenschläger)",
	"sizes": [{
	"width": 220,
	"height": 124,
	"url": "https://api.dw.com/image/42504817_301.jpg"
	},
	{
	"width": 460,
	"height": 259,
	"url": "https://api.dw.com/image/42504817_302.jpg"
	},
	{
	"width": 700,
	"height": 394,
	"url": "https://api.dw.com/image/42504817_303.jpg"
	},
	{
	"width": 940,
	"height": 529,
	"url": "https://api.dw.com/image/42504817_304.jpg"
	}
	]
	}

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

4.2.3 DIAS: Article information

During the procedure of raw data collection DIAS plans to collect article information once the article is published.

Metadata that has been identified and can be captured are:

Article Metadata	
Id	Unique identifier for the article. Used for joining the article onto other datasets
Title	Title given to the article
Path	url slug for the path pointing to this article
ImageUrls	Urls to images contained in the article
Introduction	Paragraph of text shown on smaller views of the article, for example on the front page.
Subtitle	Subtitle given to the article
PublishedDate	Date of publication
Tags	Tags given in the CMS system, category etc.
Type	An indication if the article contains video, or other type of data etc
Links	URLs pointing to this article

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

4.2.4 VRT: User viewing behavior

This dataset contains a time series describing when users have visited specific pages and how long they've been looking at the content.

User View	

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

Title	Page title
Browser	Chrome, …
Country	BE
Domain	The domain name of the document (what's in the browser bar)
Host	The reported domain (the dashboard the data goes to).
Idle	True | False (Is the person idle e.g. active in another browser tab?)
New	True | False First time visitor for the site in the last 30 days
Os	Windows | ios | android | mac | linux
PageTimer	Time to finish loading the dom.
Path	Path of the page from location.pathname
Platform	Desktop | mobile | tablet | unknown
Read	True | False (Is the person actively reading? Determined by scrolling behaviour)
Region	
Token	Temporary uuidevent's page session (regenerated when moving to another page).
Uid	The chartbeat account
User	User token.
Write	Is user currently writing? True or False
engaged_sec	Number of seconds on the page actively doing something.
internal_referrer	Filled out when document.referrer is on the same domain.
ip_address	IP address
lat	Geolocation.lat
Lng	Geolocation.lng
page_height	document.body.scrollHeight.
referrer	Referrer from document.referrer
scroll_top	window.pageYOffset or document.body.scrollTop or document.documentElement.scrollTop
time_spent	Number of seconds on the page.
user_agent	Browser user agent string

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

utc	Timestamp of event
window_height	window.innerHeight or document.body.offsetHeight.

Data pedigree

During the first phase of the project, concerns have been raised by project partners and data scientists about the methodology used for computing and capturing inferred data such as the **read**, **engaged_sec** and **referrer** fields in the above user view table. Due to the fact that this data is captured by a third party, we cannot answer questions on the exact methods used to determine these fields at the time of writing.

Before the the next status report, we intend to either query the data provider in question about their methodology, or replace the provider with a solution we have full control over.

4.2.5 Deutsche Welle: User viewing behavior

Deutsche Welle will provide data regarding user viewing behavior as stated in the chart below for content in English and French. It will be possible to add datasets related to content in further DW languages.

User View	
Article ID (OID)	Article Identification
Title	Page title
Browser	Chrome, …
Country	DE ect
Domain	www.dw.com

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

4.2.6 DIAS User viewing behavior

DIAS plans to implement a custom solution or software for handling the collection of user information. That should make it possible to generate sessions and user IDs for a period of time. Due to the amount of data this solution should be able to process we will set this up on an Azure Infrastructure. The below information should be collected:

User View	
Title	The Page Title
Browser	Safari, Chrome, etc
Country	We will need to integrate with an API in order to provide the country e.g CY,GR
Domain	www.sigmalive.com
Host	Sigmalive.com
Idle	If the page is scrolled or a mouse is moving, or not
New	True | False First time visitor for the site in the last 30 days
OS	Windows | ios | android | mac | linux
Pagetimer	Time to finish loading the dom
Path	Path of the page
Platform	Desktop | mobile | tablet | unknown

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

Internal Referrer	Filled out when document.referrer is on the same domain
Token	Temporary page session (regenerated when moving to another page).
User	User Token
IP Address	IP Address
Page Height	Scroll height information
Referrer	Referrer
User Agent	Browser user agent string
UTC	Timestamp of event
Window Height	Window inner height

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

4.3 Additional Datasets to be collected at VRT

A brief summation of additional datasets VRT is trying to collect and provide:

- The exact location of articles on the frontpage and the time series of changes made to this location.
- A time series of editorial actions taken that affect the lifecycle of an article on the news website, such as publication and republication, promotion on social networks, changes of article location.
- User's Facebook Likes.
- Datasets external to our organisation itself, for example weather data, trending topics on Twitter etc.

Regarding VRT's progress on these additional data collection tracks;

We currently have a proof-of-concept feature available representing the location (and changes to this location) of an article on the front page of VRT NWS. We are currently working on implementing a more robust version of this, along with capturing more detailed editorial actions such as updates to title and content, changes in location and presence elsewhere on the site and in the mobile app etc. This is largely an exercise in business diplomacy, due to this nature we're currently negotiating a reasonable timeframe by which this functionality can be provided.

We've ran a first successful trial capturing and processing tweets related to subjects of interest. We intend to run further trials during the 2018 local elections. The goal of this exercise will be to to capture the social buzz generated by specific events, and relate them to specific articles so that they can be used as data features.

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

4.4 Additional Datasets to be collected at Deutsche Welle

In accordance with the current Deutsche Welle data policy Deutsche Welle is not implementing further user behaviour data collection activities.

4.5 Additional Datasets to be collected at DIAS

DIAS will explore ways to provide additional datasets for CPN before the next reporting period depending on the nature of the personalization service. Among the solutions proposed are the possibility to collect and provide additional data from Sigmalive's Facebook page. These data might be able to help machine learning create correlations between users on social media in order to deliver more personalized content. The usage of this feature is closely related to the proposed scenario and solutions regarding the personalization content that we will try to deliver.

4.7 Storage formats

There are plenty of possible storage formats, but the following ones are the generally used and recommended.

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

Text

- Includes formats such as CSV, JSON etc.
- Bulky and inefficient in storage
- Extremely fast to read and parse

Apache Orc

- Row-based file format
- Works well in Hortonworks Hadoop distribution
- Allows for parallel processing of rows and columns

Apache Avro

- Row-based file format
- Good fit for Extract-Transform-Load workloads (uses all columns)
- Focused towards write operations
- Works well in combination with Apache Kafka since versioning of datasets using Kafka is implemented with Avro
- Supports schema evolution very well

Apache Parquet

- Column-based file format
- Good fit for queries where only a subset of the available columns is used
- Focused on analytical work; Write once and read or query many times
- Works well in Cloudera Hadoop distribution
- Supports schema evolution reasonably well

It's worth noting that the choice of storage format heavily depends on its intended use case. We give you some brief examples:

- Data schema will evolve continuously; Choose Avro for decent schema support
- Data is intended for heavy-duty analytics; Choose a compressed format, Parquet if typically only a subset of the available columns is used. Take notice of tooling support in the environment (Cloudera, Hortonworks, …) in which it will be used

Providing the right format for the right job

The data streaming platform in operation at VRT R&D provides one important advantage: it allows for streams of data to be replayed and reprocessed on-demand. Practically speaking, this means that the ingested raw data and its direct derivatives can be streamed to any desired format as long as they are stored in the streaming platform.

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

On the date of writing, the data retention in our streaming platform is set to six months. Any raw data within this six-month rolling window of time can be immediately streamed to any desired serialization format. If a larger rolling window is desired, we'll have to set up a two-speed system where data is serialized in one format for storage past the streaming platform's retention date, and then processed to another format if a particular use-case desires so.

Because of this, it seems canonical to serialize data to an Avro format to preserve it past the six-month rolling window. We have chosen Avro because of its great support for versioning (or schema evolution), its great performance in ETL workloads and the maturity of the format.

When the project reaches a more analytically focused phase, a six month rolling window of data can be supplied in a format of choice. If older data is required, this could be achieved by a two-speed system where the first speed is obviously the streaming layer. And the second would be to either load older data back into the streaming layer for stream processing, or following a more classical approach using something like Hadoop in a typical Lambda architecture.

4.8 Delivery of Data and Architecture at Dias

Giving access to SQL Databases directly would be ideal, instead of generating CSV files which will be difficult to manage from both sides. Another possible solution is a Web Services implementation which will expose this information to the platform.

The solution would be to setup an Azure Infrastructure. This will ensure the best performance of the solution without affecting Sigmalive's Infrastructure. This infrastructure would consist of:

- Sql Instances: 1 (storing both Article Information and User Behavior Data)
- App Services: 1 (User Behavior Application that would be executed in each page on sigmalive.com)
- Azure Queue Storage: 2 (Adding User Behavior and Article Information Data)
- Azure Functions: 2 (Collecting from Queue User and Article Information Data and importing to Sql)

Dias is also exploring other solutions in order to provide data for CPN. More information will be available before the next reporting period.

5. Conclusions

5.1 Publishing partners' data collection status

VRT currently has an initial real-time metadata streaming platform set up. This system allows us to reliably capture the usage data described elsewhere in this document in a near real-time manner with ~50ms of latency. A data science sandbox has been provisioned to allow data scientists the

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

freedom they need to flexibly work with this dataset. Specifically, this is a Jupyter notebook connected to a single Spark node in order to perform analysis over large datasets stored in Amazon S3. This can be used both to explore and gain insight into the available data in an analytical manner, and to train machine learning models on this data. Gained insights and designed blueprints or models may then be translated back to the streaming platform to evaluate models in real-time, as new data points flow in.

At the time of writing, VRT provides;

- A stream of content consumer pageviews, starting on the 26th of january 2018 and currently counting 364,798,784 rows. This stream provides a six month rolling window of pageviews, and will serialize to Avro format in case more or older data is required
- A stream of VRT NWS articles, starting on the 29th of november 2017 and currently counting ~250 unique rows.
- A stream indicating whether or not an article is present on the frontpage, beginning from january 26th 2018.

As a content partner in the CPN consortium Deutsche Welle is providing an extensive set of article-related metadata as well as a limited amount of user behavior data as stated in 4.2.5. User behavior related data from Deutsche Welle must remain confidential and available only to project partners specified in the agreement of confidence at all times.

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

5.2 Moving data collection efforts forward

As more effort is expended working with the available data, it's likely to become apparent that additional data sources are needed. We may want to know specifically how long a user has been engaging with specific content, or what elements are contained inside a media item. Therefore it's crucial that the current data collection effort is an ongoing process and can be improved upon for the full duration of the project.

This iteration process is a direct result of the natural interplay between data engineering and data science roles. Due to the nature of machine learning application development, initial proof of concepts may indicate that additional data is required.

5.3 Common hurdles and pitfalls

As insights gained from an intensive data collection effort point out, the following common pitfalls and hurdles should be avoided at all cost;

Preserving ground truth

Partners should be cautious to always store the collected raw data in an unaltered state. This allows for easy data recovery when inescapable human errors are made.

GDPR

As the application of the General Data Protection Regulation draws near, partners should be aware of its implications in the project. These include, but are not limited to;

- Querying users for permission to use specific data for a specific intent
- Respecting the users' right to be forgotten
- Respecting user data queries
- Special care should be taken when data is exchanged between partners, to make sure no policies are violated.

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

6. ANNEX

LIST OF AVAILABLE DW CONTENT IN ALL DW LANGUAGES

Numbers on 5 February 2018 on dw.com

Available data on www.dw.com

Main Languages

- English: **229.452**
  - Video: **38.922**
  - Audio: **11.044**
  - Articles: **175.506**
  - Images: **3.980**
- German: **323.697**
  - Video: **46.531**
  - Audio: **5.994**
  - Articles: **266.966**
  - Images: **4.206**

All Languages

- Amharic: **23.717**
  - Video: **5**
  - Audio: **4.836**
  - Articles: **18.830**
  - Images: **46**
- Arabic: **115.328**
  - Video: **20.323**
  - Audio: **572**
  - Articles: **82.886**
  - Images: **2.602**
- Bengali: **46.261**
  - Video: **1.387**
  - Audio: **23**
  - Articles: **44.117**
  - Images: **2.734**
- Bosnian: **63.890**
  - Video: **1.434**
  - Audio: **0**
  - Articles: **62.143**
  - Images: **313**
- Bulgarian: **40.351**
  - Video: **592**
  - Audio: **0**
  - Articles: **39.578**
  - Images: **181**
- Chinese (simplified): **121.196**
  - Video: **1.241**
  - Audio: **636**
  - Articles: **117.168**
  - Images: **2.151**
- Chinese (traditional): **121.196**
  - Video: **1.241**
  - Audio: **636**
  - Articles: **117.168**
  - Images: **2.151**
- Croatian: **42.320**
  - Video: **1.567**

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

    - Audio: **115**
    - Articles: **40.389**
    - Images: **249**
- Dari: **27.286**
    - Video: **320**
    - Audio: **456**
    - Articles: **25.140**
    - Images: **1.370**
- English: **229.452**
    - Video: **38.922**
    - Audio: **11.044**
    - Articles: **175.506**
    - Images: **3.980**
- French: **43.598**
    - Video: **251**
    - Audio: **3.816**
    - Articles: **39.436**
    - Images: **95**
- German: **323.697**
    - Video: **46.531**
    - Audio: **5.994**
    - Articles: **266.966**
    - Images: **4.206**
- Greek: **28.008**
    - Video: **18**
    - Audio: **28**
    - Articles: **27.948**
    - Images: **11**
- Hausa: **30.630**
    - Video: **240**
    - Audio: **4.563**
    - Articles: **25.692**
    - Images: **135**
- Hindi: **50.822**
    - Video: **1.517**
    - Audio: **35**
    - Articles: **45.858**
    - Images: **3.412**
- Indonesian (Bahasa Indonesia): **51.431**
    - Video: **1.373**
    - Audio: **0**
    - Articles: **47.448**
    - Images: **2.610**
- Kiswahili: **47.408**
    - Video: **554**
    - Audio: **4.447**
    - Articles: **42.318**
    - Images: **89**
- Macedonian: **64.639**
    - Video: **1.691**
    - Audio: **0**
    - Articles: **62.634**
    - Images: **314**
- Pashto: **25.892**
    - Video: **391**
    - Audio: **541**
    - Articles: **22.975**
    - Images: **1.985**
- Persian/Farsi: **110.356**
    - Video: **1.657**
    - Audio: **1.314**
    - Articles: **102.646**
    - Images: **4.739**
- Polish: **40.518**
    - Video: **1.521**
    - Audio: **55**
    - Articles: **38.315**

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

    - Images: **627**
- Portuguese for Africa: **20.786**
    - Video: **623**
    - Audio: **4.069**
    - Articles: **15.772**
    - Images: **322**
- Portuguese for Brazil: **63.846**
    - Video: **4.202**
    - Audio: **1**
    - Articles: **58.218**
    - Images: **1.425**
- Romanian: **39.527**
    - Video: **354**
    - Audio: **0**
    - Articles: **38.893**
    - Images: **280**
- Russian: **188.786**
    - Video: **8.000**
    - Audio: **96**
    - Articles: **180.453**
    - Images: **233**
- Serbian: **49.723**
    - Video: **1.849**
    - Audio: **80**
    - Articles: **47.352**
    - Images: **442**
- Spanish: **149.263**
    - Video: **31.497**
    - Audio: **8**
    - Articles: **115.640**
    - Images: **2.118**
- Turkish: **80.844**
    - Video: **3.949**
    - Audio: **4.634**
    - Articles: **71.209**
    - Images: **1.052**
- Ukrainian: **76.163**
    - Video: **5.351**
    - Audio: **0**
    - Articles: **70.150**
    - Images: **662**
- Urdu: **54.546**
    - Video: **1.363**
    - Audio: **1.357**
    - Articles: **50.833**
    - Images: **993**

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