Grant Agreement No.: 761488

D1.3: Innovative CPN Components

This deliverable describes the innovative components for personalisation.

Projectcpn.eu

Work package	WP 1
Task	T1.3
Due date	28/02/2018
Submission date	28/02/2018
Deliverable lead	VRT
Version	1.0
Authors	Jens Van Lier (VRT), Olga Kisselmann (DW), Hans Dreesen (VRT), Michele Nati (DCat), Matthias De Vriendt (VRT), Joris Mattheijssens (VRT), Tilman Wagner (DW)
Reviewers	Nikos Saris (ATC)
Keywords	CPN, personalisation, innovative, personal data

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	22/02/2017	Initial version	Jens Van Lier (VRT), Olga Kisselmann (DW), Hans Dreesen (VRT), Michele Nati (DCat)
V1.0	28/02/2017	Final version	Hans Dreesen (VRT), Michele Nati (DCat), Bart Janssens (VRT), Olga Kisselmann (DW)

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.

This document reflects only the authors' views, and the Commission is not responsible for any use that may be made of the information it contains.

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~~

EXECUTIVE SUMMARY

This deliverable describes the innovative components that can add value to CPN's technical framework.

It explains why, in order to help publishers personalise their content, it is important to go beyond the state of the art and to introduce novel techniques and approaches.

It explains how our research, workshops, surveys and meetings with industry experts directed our attention to the following components:

- **Newsbots and smart speakers**
- **External and highly contextualized datasets**
- **Multi-layered personalisation**

The document describes these components in detail and concludes with an outlook on how they could be used in the CPN project.

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

LIST OF FIGURES

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

ABBREVIATIONS

GDPR	General Data Protection Regulation
DW	Deutsche Welle
API	Application Programming Interface
T&C	Terms & Conditions
PDR	Personal Data Receipt
NLP	Natural Language Processing
LIWC	Linguistic Inquiry and Word Count

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

1 INTRODUCTION

If we want to help publishers reach and inform their audiences better, we have to look beyond current personalisation techniques. It has been proven that when used in a careless way, personalisation might create biases and erode trust. We will prevent those negative side effects by using the highest standards in privacy and transparency, but also by adding value for both the end user and the publishers. The components we propose will better align with the users' context and personalities. They will take in account external factors, but also align with innate characteristics of the individual users. For the publishers this means a higher engagement with their content and a possible increase in audience size.

To do this, we selected the techniques from our research that are new to the area of news publishing or have great potential. The design thinking workshops yielded interesting ideas and concepts that were enriched with the results from our user surveys. A panel of industry experts informed us further, and confirmed our choices.

News bots, smart speakers and digital assistants will allow to make the user experience of news consumption as seamless as possible. They rapidly gain in popularity and offer a whole new paradigm for personalisation.

External data sources will add intelligence to personalisation on an individual level. Taking a user's context into account while increasing engagement.

A dedicated focus on **trust and transparency** will avoid suspicion and privacy fatigue (Choi, Park, & Jung, 2018). Beyond what is required by GDPR, there is an opportunity to add value for users in exchange for voluntary disclosure of extra data. This might also create new revenue opportunities.

Lastly, we propose to open up personalisation to multiple levels and use sophisticated techniques to take the personal traits of users into account. We call this approach **multi-level personalisation**. This will allow the system to create parallel article versions that will be better adapted to the individual user's characteristics instead of today's *one size fits all* approach.

2 INNOVATIVE COMPONENTS

As mentioned in the introduction, below we will discuss the following domains into more detail.

- Newsbots and smart speakers
- External and highly contextualized datasets
- Security, trust, control and transparency
- Multi-layered personalisation

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

2.1 NEWSBOTS AND SMART SPEAKERS

We decided to include text- and voice-based conversational applications into our vision of the CPN platform, as this technology is prognosed to grow significantly in the near future and has the potential to provide a user-friendly and intuitive interface for CPN. During our research activities for initial CPN requirements, which included co-creation session with the target group and an extensive online survey, we got a fine-grained and comprehensive picture of users' preferences and expectations for a news personalisation platform.

While users want to be in control and actively influence their news recommendations, they also wish for a more engaging, intuitive and seamless way to receive personalised news. The emotional aspects of news consumption play an essential role in the user's perception of new technologies. Developments in the area of text- and voice-based conversational bots could tackle this challenge and provide a viable addition to our platform. By engaging with a chatbot, CPN users could pull content, state opinions and communicate their current emotional state or fine-tune the recommendation settings.

In recent years, chatbots have become increasingly popular. However, the technology for chatbots is not new, as first chatbots have been developed in the 1950s[1]. The bot Eliza, created by Joseph Weizenbaum in 1966 was one of the first humanlike chatbots. Eliza was programmed to mimic a dialogue based on theories of the Rogerian speech therapy, with an emphasis on an emphatic, but passive communication style. Eliza can ask leading questions, similar to a Rogerian speech therapist, scan for keywords and ask further suitable questions from an extensive database. Since Eliza, there have been many attempts to create humanlike chatbots, leading to extensive research and an annual contest for the most humanlike conversational program, the Loebner Prize[2].

Today's chatbots are able to operate with advanced language processing and machine-learning technologies, but still function in a similar way to Eliza: they are programmed to match specific keywords/input triggers provided by the user with a predefined dataset, usually on a closed domain, where they specialise on certain (conversational) topics or tasks only. More sophisticated bots can answer logically and can communicate in open domains, still with particular limitations. No chatbot so far has been able to pass the initial test for artificial intelligence formulated by Alan Turing in 1950, i.e. "fool" a user over an extended period of time, that it is human.

While apps are still dominating consumers' digital habits, with 2,2 million apps for iOS and 2,8 million apps for Android currently available[3], consumers are reluctant to try new apps[4]. The majority of users engage with just a few applications, and messaging apps are especially popular[5]. WhatsApp hit 1,5 billion active users per month in December 2017, while Facebook Messenger counted for an additional 1,2 billion active users in April of the same year[6].

[1] Shum, He &Li: 2018 https://arxiv.org/abs/1801.01957

[2] http://www.aisb.org.uk/events/loebner-prize

[3] https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores

[4] https://techcrunch.com/2017/08/25/majority-of-u-s-consumers-still-download-zero-apps-per-month-says-comscore

[5] https://www.statista.com/statistics/260819/number-of-monthly-active-whatsapp-users

[6] https://www.statista.com/statistics/417295/facebook-messenger-monthly-active-users

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

Since 2015, when the messenger service Telegram opened its platform for third-party bots, the messenger quickly became populated with conversational bots of all kinds. Among them, many media and news-related applications and major messenger services like Slack, Facebook Messenger and WeChat. Various other media and tech companies followed the trend of giving third-party providers access to their platform or even providing convenient building blocks for customised chatbot development. In April 2017, more than 100,000 chatbots have been available for Facebook Messenger alone[7]. The potential global annual revenue generated by chatbot transactions is estimated to account for up to USD 32 billion[8].

Voice-based digital assistants are a more recent technology. They have the potential to change product design and our conceptualisation of human-machine interactions at large. Voice-based applications can handle various tasks and provide a better user experience for visually impaired users. The market leader for voice-based smart digital devices, Amazon's Echo, with its conversational interface Alexa, has over 1400 different functions. Voice-based digital assistants and devices are becoming increasingly popular. According to the German Federal Association for the Digital Economy, 56% of Germans have already, at least once, interacted with a voice-based digital assistant[9]. The market research institute Juniper estimates that by 2020, 70 million US households will have at least one smart speaker in their home[10].

Current developments in text and voice-based digital assistants go in two directions. On the one hand towards more intelligent bots that mimic a humanlike conversation, on the other hand towards so-called "dumb bots"[11] that just follow certain commands and provide output for specialised tasks or information queries. Examples are Microsoft’s conversational bot for the Chinese market Xiaoice[12], which is made with the intent to imitate human interaction or Amazon’s Echo, that follows the “dumb” approach, where a long conversation is not necessary for the device to perform specific tasks.

Another significant development lies in the area of emotion detection. The interaction with bots is often described as chunky and emotionless. To convince users, bots have to learn to read the emotional underlinings of conversations and be able to provide information and entertainment in accordance with the emotional state of the user. One prominent example of a research project that tackles this challenge is the ECM[13] (Emotional Chatting Machine) project. ECM is a chatbot, which is not only able to engage with the user by providing a factually coherent conversation but also to spike this conversation with eventual expressions of emotions, corresponding to the emotional state of the user, which is detected by analysing keywords, expressions and speech patterns. To teach the machine about human emotions and suitable responses an “emotion classifying” algorithm was developed by the research team. The “algorithm” learned to detect emotion from 23,000 posts taken from the Chinese social media site Weibo[14]. The posts have been manually classified by humans into categories like “happy”, “enraged”, “sad” and so on.

[7] https://venturebeat.com/2017/04/18/facebook-messenger-hits-100000-bots

[8] Zumstein, 2017: https://www.researchgate.net/publication/322855718_Chatbots_-_An_Interactive_Technology_for_Personalized_Communication_Transactions_and_Services

[9] https://www.bvdw.org/der-bvdw/news/detail/artikel/bvdw-studie-mehrheit-nutzt-digitale-sprachassistenten

[10] https://techcrunch.com/2017/08/25/majority-of-u-s-consumers-still-download-zero-apps-per-month-says-comscore

[11] https://venturebeat.com/2017/02/03/screw-the-turing-test-chatbots-dont-need-to-act-human/

[12] http://nautil.us/issue/33/attraction/your-next-new-best-friend-might-be-a-robot

[13] https://github.com/loadder/ECM-tf

[14] https://www.weibo.com

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

The emotion classifier was then used to tag millions of social media interactions according to emotional content. After these steps, the bot could be trained to detect keywords and patterns in the user's textual input and provide corresponding reactions in a written conversation[15].

Quartz Media, one of the industry leaders in the area of digital application and bot development, is trying to add an individual touch by delivering news to users via its Quartzy newsbot[16] that first has been curated and discussed by (human) editors over a messenger app. The bot not only provides the news stories but also elements of the editor's discussion like the emoji used as reactions. This way the application emulates a more natural conversation with the user. Chatbots are easy to deploy, integrate well with new platforms and existing services and provide publishers with convenient tools to open new distribution channels on existing messaging platforms or stand-alone applications. Quartz Media uses both strategies to reach customers.

While there are various examples of how news publishers can use conversational applications, including TechCrunch, The Wall Street Journal, CNN, The Guardian but also DW's own newsbot application, we also examined chatbots used in other industries to outline a suitable use case of text- or voice-based conversational applications for CPN. One particularly interesting application is the conversational bot Lara, created by the dating site Match.com[17]. The bot is deployed to a messaging platform, asks users questions and then creates a corresponding profile for future recommendations. To complete this process, the users don’t need to leave the messenger platform or open Match.com‘s main application.

For CPN a similar approach could be applied to engage with users over time and create or update a news-preference profile through conversation, games and quizzes. Such a use case would correspond to the preferences and expectations articulated by test users during our requirements evaluation research activities, as seen in D1.1. Results of the DW user survey indicate that a majority of respondents is willing to interact with a news recommendation application on a daily basis and prefer to actively influence their recommendations instead of passively providing data via third-party APIs and behaviour tracking. An intelligent conversation application, able to ask interesting questions that go beyond the usual preselection of content categories was also mentioned by users in the co-creation sessions conducted by DIAS.

To develop a suitable solution for CPN, we will have to take into consideration the challenging aspects of text- and voice-based bot applications. Among them, the users’ limited attention span and possibly quick frustration with chatbots, limitation of natural language processing technologies, as well as the complexity to accommodate highly individual language usage and language differences among our international audience (without making it too tiresome to learn to communicate with the system).

2.2 EXTERNAL AND HIGHLY CONTEXTUALISED DATASETS

How can the use of external data sources like traffic information and weather data contribute to better personalisation? The usage of sensors and the data coming from devices like smartwatches can add a new stream of very personalised and contextualized data to the data already available.

[15] Zhou et al, 2017: https://arxiv.org/abs/1704.01074

[16] https://quartzy.qz.com/1119980/meet-quartzys-cultural-companion-bot

[17] http://www.thedrum.com/news/2017/04/19/matchcom-launches-lara-dating-chatbot-help-people-find-love-using-ai

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

Another source is data from social media platforms, available through open APIs and Oauth authentication, based on user consent. This will allow to better understand user preferences and to avoid duplicating and recreating that data from scratch. It will also satisfy the data minimisation principle of Privacy by Design. It will, however, require to clearly and transparently explain to users how their data will be used, or how it will alter their preferences (see Section 2.3). Not all social media platforms currently accommodate linking their users' profiles to other platforms, so CPN will have to experiment first with the few available ones in order to create opportunities that are valuable for all parties.

Integrating a Facebook login currently allows apps and services to import Facebook users' profiles, including preferences, if they allow it. This model has been successful for Facebook because it removes registration fatigue and apps can leverage the Facebook advertising network for additional revenue streams. However, a similar model hasn't been as successful for other popular social media. For instance, Spotify's APIs allow to integrate the platform with other services, but not to export users' profiles. The GDPR will require the guarantee of an individual's right to data portability, and it is expected that businesses will comply with that requirement to avoid losing customers. The right set of incentives and the return of value derived from such an integration still need to be identified. CPN will investigate the opportunities to return value to those platforms. It might include importing a new set of user data generated by CPN, to those platforms, to allow them to also better personalise their services (e.g. to recommend music based on the user's mood or activity).

2.2.1 LEARNING A USER'S CONTEXT

Each user's news and content consumption is highly influenced by his or her context. A user's context is defined by all factors that define a person at a given time. These factors can both be internal, e.g.:

- What he or she is doing at a given time
- A person's current emotions and mood
- The knowledge a person has regarding a certain subject

Or they can be external in nature, e.g.:

- The importance of the subject with regards to events that society as a whole deems important or unimportant. For example, is the topic trending or not? E.g. content regarding the migrant crisis in Europe, ongoing elections, etc.
- Time of year and meteorological conditions. Is it a bright summer day? Or is the person stuck in a traffic jam during a particularly harsh winter day?
- Are there currently any ongoing high-profile political events such as elections?

The importance of this type of personal context is non-trivial. When we're dealing with machine-learning algorithms we could imagine a system that attempts to output a prediction in two steps, as follows:

Step 1: For any given user A, we predict the probability P of their receptiveness to (news-related) content given their context B.

Step 2: If the user is found to be receptive to (news-related) content (i.e. P is high), which of the available items is most interesting for user A in context B.

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

Step 1 could roughly be rephrased as "How likely is a user to engage with news-related content depending on a given context?". A naive example of such a prediction could be as follows.

For user John Doe:

Known user context	***Prediction for probability P that user is receptive to engage with news-related content***
During breakfast, while preparing to go to work	60%
Commuting to work using public transport	83%
While working at their desk-job	17%
During lunch break	72%
During free time in the evening, while at home	25%

If such a prediction was made for John Doe, then perhaps it wouldn't be wise to push notifications of recommended content while he or she is busy at work. Perhaps it would be better to wait and push these only when we know he or she is now on lunch break and much more inclined to engage with our content.

Note that this example of user context is mainly based on a user's activity and daily schedule. Predictions could be made on a wide variety of inputs, such as a user's emotional mood, or external factors such as trending topics.

The complexity of such a prediction lies in the detail that the input for the prediction, a user's context, will never be available as an undeniable fact. Instead, it's inferred knowledge. For example, using data from GPS and motion sensors available in common mobile phones we detect that a user is moving at a very high speed along a motorway connecting two metropolitan areas. While these are actual facts, inferred knowledge could include:

- Based on speed and location we infer the user is driving a car
- Based on the time of day and the user's daily routine we infer he or she is commuting to work
- Based on regular and prolonged reductions in speed and available meteorological data, we infer the user is stuck in a traffic jam

2.2.2 DATA SOURCES

To accurately infer knowledge about any given user, we must first have a steady supply of facts about this user and his environment. There are many ways by which we may choose to do so.

Geolocation, commodity hardware and sensors

Most mobile phones (or smartphones) in circulation today are loaded with hardware sensors and system services that can be used to gain key insights into a user's context. Many of these devices

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

are outfitted with GPS chips and accelerometers. While these can be accessed individually, it is often more interesting to integrate with the available services that are built on top of these sensors. It's worth looking into

- the Core Motion service for iOS, developed by Apple
- the Motion Sensors and Detected Activity APIs for Android

Using these services, it becomes trivial to learn a user's location, means of travel, speed etc.

Smartwatches and fitness APIs

Smartwatches offer developers and manufacturers detailed insight into a user's sports activity, heart rate and even sleeping activity. While this type of device has not yet found its way to a large part of the population yet, services such as the Google Fit API attempt to compensate by aggregating more common sensor data into distilled fitness and health metrics.

Swarm applications

Swarm or crowdsourcing applications have recently found their way to mainstream usage. Think of Waze, the Google Local Guides Initiative and Strava. As more companies value this data very highly, the realisation grows that aggregated datasets can achieve much more than a single user's data ever could. This opens up possibilities for new types of applications and services.

Take Waze as an example. When you know not only the context of each of your active users but this number of users grows large enough to represent a large part of all participants in a certain traffic situation, then you may infer additional bits of useful knowledge. In the case of Waze: is a given route jammed or not? What are the average throughput times for any given route? What would be the fastest route from A to B, keeping in mind typical throughput times during this time of day?

The rise of the personal assistant

As devices like the Amazon Echo and Google Home find their way into the mainstream households, new types of interactions and data collection are possible. How many family members are at home? Are they having dinner or watching television? What's the current mood in the household? Are people laughing or arguing?

Social media

Social media form the easiest and most effective way to gain insight into trending topics and public opinion. Trending hashtags often offer the additional quality of being much more real-time than investigative journalism. When certain subjects gain traction and begin peaking, this can often be detected long before the first articles on the subject appear on popular news outlets.

At this point, news outlets often integrate social media into their reporting, thus creating a direct link between popular opinion on a topic and the relevant news items.

Additionally, an aggregation of tweets surrounding a certain hashtag often provide a fairly complete story of breaking events (such as terrorist attacks) long before these can be distilled into comprehensive articles.

Facebook's emoji-responses provide a fast and easy, albeit qualitatively abysmal, way to sample public opinion regarding a certain topic.

Public and open source datasets

- Real-time traffic information on congestion, accidents etc.
- Open meteorological data
- The national train company's real-time API

2.2.3 PIECING TOGETHER FACTS TO INFER CONTEXT

The table below offers a few basic examples of user context, and the data sources from which it could be effectively derived. In order to determine whether or not a certain user context can be derived from a certain data source by a machine-learning algorithm, a good question to ask would be: "If a domain expert was presented with this data, would he be able to accurately infer the context?"

Specific user context	*Data sources*
Is user at home, at work or on the go?	Location services, patterns in time series and aggregates
Means of transportation (car, train, bicycle, etc.)	Location services and motion sensors
Is user sleeping or out on a Saturday night bar crawl?	Motion sensors

Deriving personal traits from social media activity

A useful proxy to gather personal traits of users is their activity on social media. Researcher Michal Kosinski demonstrated in 2013 that personal traits such as sexual orientation, ethnicity, religious and political views and use of addictive substances, amongst others, could be accurately predicted by only looking at an individual's Facebook likes[18].

Furthermore, a framework for detecting a person's topic preferences given their activity on Twitter is available from the computer science and political departments of the University of Rochester[19].

2.3 SECURITY, TRUST, CONTROL AND TRANSPARENCY

In times of large data hacks, shady deals between web portals and data brokers and intransparency regarding information shared on social networks, users are getting very wary about their online personal data. At the same time people don't want to stop using social media

[18] *Private traits and attributes are predictable from digital records of human behavior* by M. Kosinski, D. Stillwell, T. Graepel, Proceedings of the National Academy of Sciences (PNAS), 2013.

[19] *Catching Fire via 'Likes': Inferring Topic Preferences of Trump Followers on Twitter* by Yu Wang, Jiebo Luo, Richard Niemi, Yuncheng Li, Tianran Hu

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

or quit the comfort of using online services such as web shops and online banking. This creates a difficult starting point for good personalisation. To build meaningful user profiles and provide relevant content, personalisation algorithms are very dependant on the availability of this kind of data and the willingness of users to share their interests and habits.

This is also clearly visible in the user survey, that DW conducted among its target audiences. DW sent out two questionnaires targeting English speakers and Spanish speakers as two of the largest audience groups. While 77% (English) and ~83,5% (Spanish) of all participants in the survey stated that they were willing to share data with an application in order to get a more personalised offer, the majority in both surveys also clearly stated that they wanted to be in control (~84,5% English / ~85,7% Spanish) and in the know (~85% English/ ~80,7% Spanish) about the usage of their data.

This implies that any new application or service, trying to serve a personalised offer of any kind has to build up trust among its users from the first moment on. Looking at the current situation, we think there are three main aspects to take into consideration to build this level of trust:

a) **Security:** The project must convince users that their data is safe from external access (data hacks) and not to be shared/sold to anyone without their explicit consent (data brokerage)
b) **Transparency:** Users must always be able to see what data about them is stored and what it is used for in an understandable way.
c) **Control:** Users must also be able to make changes to their data, e.g. withdraw their consent for its use, change it or even delete data, in parts or completely in an easy fashion.

While security should be a standard element of all applications dealing with user data, transparency and control over the stored data are not always in a company's interest, as the examples of several big social networks show. Their business model is the use of such data and also the intransparency of their algorithms as their main USP (e.g. Facebook). However, we believe that by keeping the usage transparent and allowing the user for more control over their data, a much closer relationship can be formed.

Furthermore, there are different initiatives and legal frameworks currently underway or in the making, that are aiming to give users a more secure basis, like the GDPR, regulating the use their data.

In fact, on the one hand, GDPR requires organisations, large and small, to perform a greater degree of due diligence when dealing with a customer’s personal data, achieved by carefully reviewing their processes and including Privacy and Security by Design principles (and in some case Data Protection Impact Assessments). This is being instituted to avoid large fines handed out due to possible data breaches. This will ensure higher-level of security when dealing with personal data, currently demanded by users.

On the other hand, the higher demand for transparency and granular consent, supported by the GDPR and demanded by users to organisations before accessing their data, now offers opportunities to create new channels and regain users’ trust.

GDPR Article 4 states that consent should be freely given, unambiguous as well as specific to the purpose, while Article 7 requires that proof of such consent should be maintained by both parties, the *data subject* and the *data controller*.

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

Transparency and control are at the core of Article 12-14 that requires *data controllers* (e.g., the organisation deciding the purpose for collecting and using personal data as part of a named service offer) to provide *a fair processing information notice on how the subject's personal data will be collected and used*. This aims to increase transparency over how organisations use the personal data they collect. Similarly, Articles 15-19 demand more user control over their data. Article 15 outlines the *right of access* by the *data subject*, with Articles 17-19 regulating the rights of the individual, including the *right to be removed from databases,* and consequently any personal link to his/her data be removed upon request.

For CPN, it is clear how a proper implementation of Articles 12-14 will offer the chance to satisfy the users' need for transparency and control and to regain trust, provided that the needs of users (always more mobile) also looking for a better user experience are clear, and that simple privacy statements are taken into account. Although recommendations on how to implement specific GDPR articles have started to emerge, it will be hard to satisfy these requirements with a one-size-fits-all solution.

A possible compliant solution that the project will explore and that provides the transparency required by GDPR and by users are: The Personal Data Receipts (PDRs)[20].

PDRs are a human-readable record summarising in a simple and clear way what personal data an organisation is collecting about an individual, for what purpose, how they are stored, for how long and if any third party sharing is allowed.

PDRs have been co-designed with end-users. Despite the amount of details provided by Privacy Policies and T&Cs, the interviewed groups deemed relevant a summary of the following information:

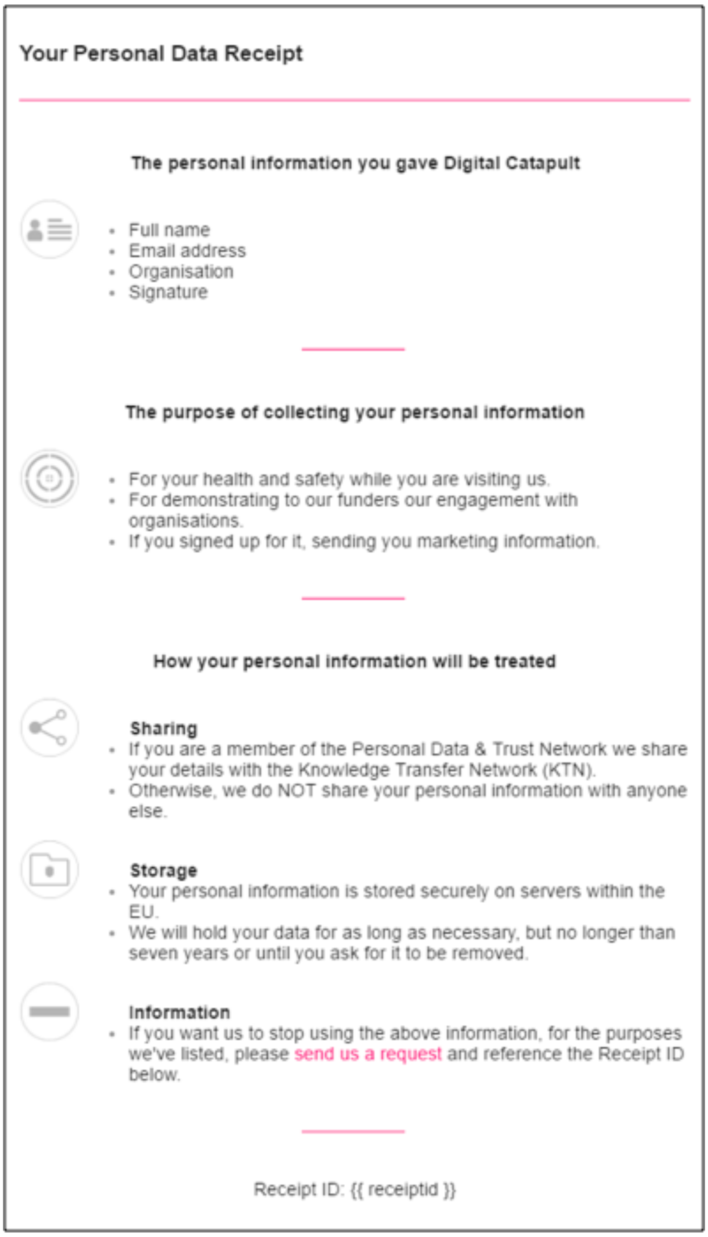
1. The category of *personal information* the organisation collects to provide a subscribed service.
2. The purpose of collecting the *personal information*, with particular emphasis on envisioned third-party sharing.
3. The where, how and the length of time the *personal information* is stored.
4. The contact details of the *Data Controller* (i.e., who sets the purpose for the personal data collection) to easily flag the request for removal of shared personal information.

Sampled groups welcomed the use of icons, but only if supported by simple, non-technical, plain text, which was considered as the main requirement for a meaningful Personal Data Receipt.

The figure below shows the implementation of a Personal Data Receipt, including the categories of information as per the above bullet points.

[20] https://www.digitalcatapultcentre.org.uk/project/pd-receipt/

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



Your Personal Data Receipt

The personal information you gave Digital Catapult

- Full name
- Email address
- Organisation
- Signature

The purpose of collecting your personal information

- For your health and safety while you are visiting us.
- For demonstrating to our funders our engagement with organisations.
- If you signed up for it, sending you marketing information.

How your personal information will be treated

Sharing
- If you are a member of the Personal Data & Trust Network we share your details with the Knowledge Transfer Network (KTN).
- Otherwise, we do NOT share your personal information with anyone else.

Storage
- Your personal information is stored securely on servers within the EU.
- We will hold your data for as long as necessary, but no longer than seven years or until you ask for it to be removed.

Information
- If you want us to stop using the above information, for the purposes we've listed, please send us a request and reference the Receipt ID below.

Receipt ID: {{ receiptid }}

Figure 1: a Personal Data Receipt

For simplicity of implementation and delivery, PDRs can be issued as mobile-friendly emails when a new digital service is joined.

PDRs answer the need for transparency and control, and clearly communicate the level of security undertaken by a company dealing with personal data.

By making use of icons, plain simple text and hyperlinks to additional tools for managing user consent and other digital rights (like data erasure), PDRs provide a GDPR compliant solution to increase the knowledge and control users have over their personal data.

By communicating in a transparent way, Personal Data Receipts build a compliant new channel that increases consumers' trust, (perceived) feeling of transparency and control. This in contrast to the complexity of legal terms in Privacy Policies and Terms and Conditions, which only 20% of

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

European citizens admit to read in full[21]. According to the Mobile Ecosystem Forum Consumers Trust 217 report[22], over 60% of consumers value transparency as a metric to trust an organization that handles their data, and to share even more with it.

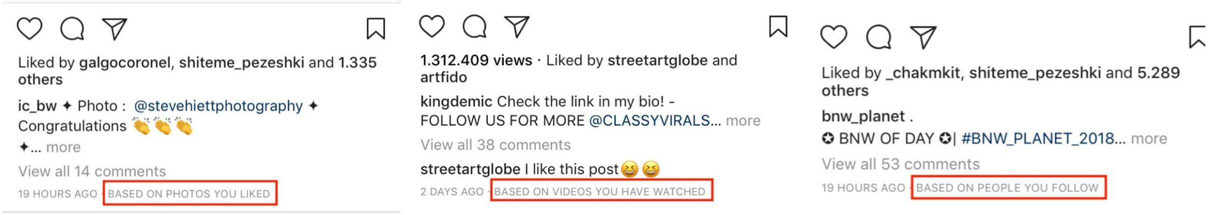
By increasing trust through transparency, the Personal Data Receipts could be also a tool to allow readers to connect more of their accounts and exposing their data to CPN through open APIs. As a Personal Data Receipt is issued as an email when a reader joins CPN, the email address could also be used to discover other social media the user is registered on. In case those social media allow an export of the user's profile, a customized email receipt could be sent to the users to explain how and for what purpose they could connect those accounts to CPN. We will expand current PDRs with a section, "Connect more accounts", specifically designed for this purpose.

Additionally, as Personal Data Receipts will be stored in the blockchain. making them non-repudiable, by either data subject and in particular data controllers, this shall increase also the perceived level of control of the users on their data, with consequent increase of trust.

Finally, the GDPR and the users' demand for transparency don't limit for what purpose and in which way personal data is used at the time a new service is joined (a requirement covered by PDRs). It also includes a more dynamic explanation on how algorithms generate recommendations based on the collected data (GDPR Article 15 Right of Access and 22 Automated decision making). Although this mainly refers to algorithms and automated decisions that might generate legal effects concerning the users, it was deemed important also by CPN in the light of achieving greater transparency.

Following this last requirement, according to the results from the user workshops and the surveys we conducted, it is clear that users are interested in several levels of transparency and control. What we envision is a solution that gives users different ways of seeing and controlling personalisation. Let's take a look at three explicit examples of how this is currently done and what CPN could do differently.

Users want to know how and for what purpose their data is used. In particular this means they want to understand how, i.e. on what grounds, recommendations are made. But they also want to see *why* a recommendation is made. Instagram has started doing this by adding an extra text line to recommendations in its app. But this is still a very basic way of giving people more insight into their algorithm (see Figure 2). There is still no way to see, for example, which people I follow that are also connected to a new recommendation.



Liked by galgocoronel, shiteme_pezeshki and 1.335 others

ic_bw ✦ Photo : @stevehiettphotography ✦ Congratulations 👏👏👏 ✦... more

View all 14 comments

19 HOURS AGO · BASED ON PHOTOS YOU LIKED

1.312.409 views · Liked by streetartglobe and artfido

kingdemic Check the link in my bio! - FOLLOW US FOR MORE @CLASSYVIRALS... more

View all 38 comments

streetartglobe I like this post😆😆

2 DAYS AGO · BASED ON VIDEOS YOU HAVE WATCHED

Liked by _chakmkit, shiteme_pezeshki and 5.289 others

bnw_planet .

✪ BNW OF DAY ✪| #BNW_PLANET_2018... more

View all 53 comments

19 HOURS AGO · BASED ON PEOPLE YOU FOLLOW

Figure 2: Instagram's basic way of providing insights into recommendations

[21] Data Protection Eurobarometer: https://www.scribd.com/document/357086983/Factsheet-Data-Protection-Eurobarometer-240615-En

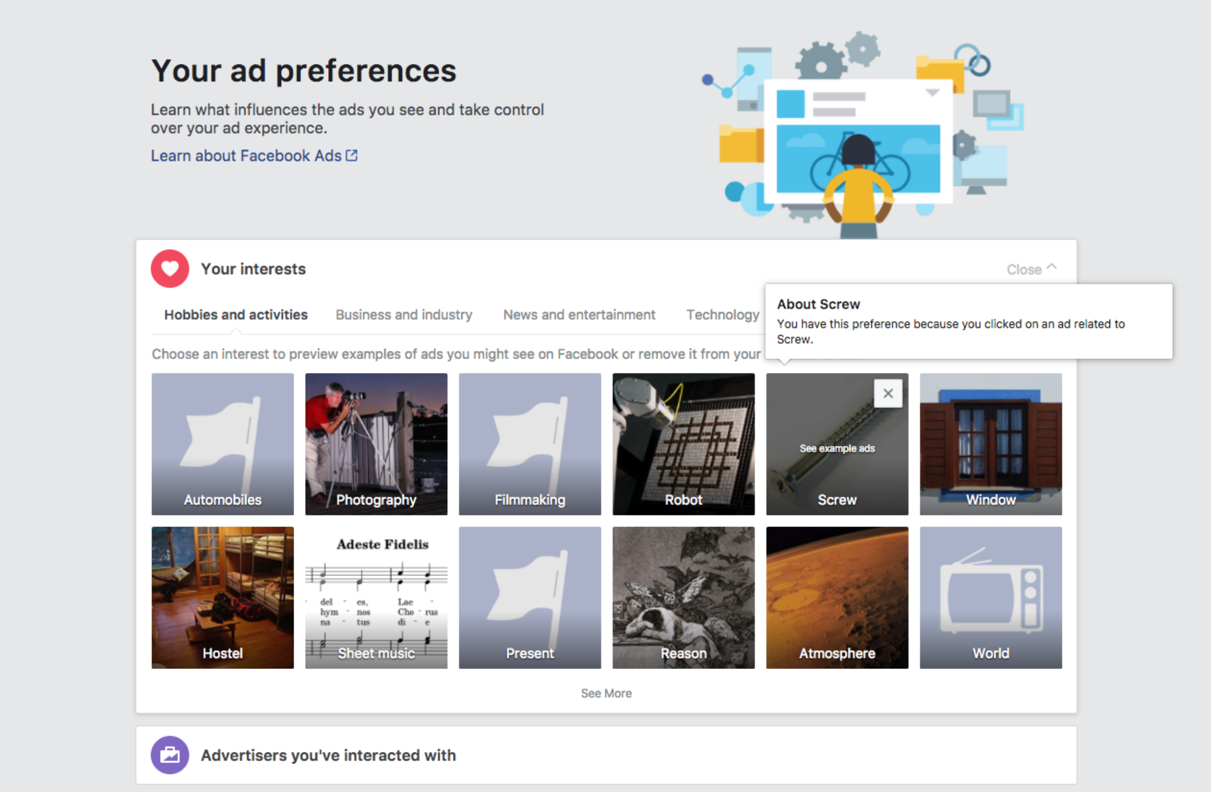
[22] Global Consumer Trust Report 2017: https://mobileecosystemforum.com/programmes/consumer-trust/global-consumer-trust-survey-2017/

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

Another example of this is YouTube that provides no information[23] regarding the choice of recommendations (the videos suggested after watching a clip on the network). In some cases, it is based on what you have watched before. But often the indication is not obvious, and there are no explanations provided by the platform, which has been highly criticized.

This is why a solution in CPN should make it clear to the user why a certain item is shown to them, directly in correlation with the item. This will need some experimenting and several iterations to get it right, but it should help increase the trust and usability of the solution.

What's also missing from most personalisation offers in the media, is a way to actively change the underlying data, influencing what people are shown. Facebook has come forward with a certain control over this, by allowing users to see and change what they like according to the network. But this is only in relation to the advertisements on the platform, not the content. And it is again limited in a way that Facebook still doesn't show how it handles this data.



Your ad preferences

Learn what influences the ads you see and take control over your ad experience.

Learn about Facebook Ads

Your interests

Close

Hobbies and activities | Business and industry | News and entertainment | Technology

About Screw

You have this preference because you clicked on an ad related to Screw.

Choose an interest to preview examples of ads you might see on Facebook or remove it from your

See More

Advertisers you've interacted with

Figure 3: Reviewing and changing ad preferences on Facebook

Particularly for the distribution of news, we think that it is necessary to make this more clear to users and give them more control. This is why we envision a dashboard that gives users the possibility to set certain criteria themselves, allowing the system to adjust them based on their behaviour, but also to overwrite these changes again or even turn them off.

[23] Deep Neural Networks for YouTube Recommendations: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf

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

Since compliance with the GDPR is mandatory for CPN, there is also the issue of allowing the user to download all data stored about him/herself in a system. This is something that is currently being offered by networks like Facebook and Twitter as well (since they also have to follow European data laws). But again, it lacks transparency and usefulness. Complying with GDPR, a possible way to make use of this is to allow users not only to download his/her dataset but to also upload it to a new system. By incorporating this kind of solution in CPN, the system could solve two things:

1. Giving users an easy to use overview of the data coming from other networks, which don't offer this kind of interface.
2. Use the user profile data already existing for its own recommendations, skipping the slow process of building a new profile for useful recommendations, given the users' consent.

While the overall aim of this is to build a solution that users are willing to entrust their data to, CPN will of course have to experiment and test which of these solutions are feasible on a technical level and to what extent. But eventually, this could lead to a better recommendation engine, that users are genuinely willing to use, knowing that they have full control over their data and that they leave it in good hands.

2.4 MULTI-LAYERED PERSONALISATION

Digital news content is generally created in a *one size fits all* fashion. Each published news item, be it an online article or web video, exists as one version only. At the same time, its intended audience is considered to be a uniform, homogenous group of people. That audience is usually drawn to (or even defined by) the publisher's brand, tone, style and levels of seriousness and complexity.

There might be a big opportunity if we refine those parameters: creating multiple versions of each news item and serving them to smaller, more accurately targeted sub-audiences by using more sophisticated ways of personalisation.

Inspired by one of the user journeys from the first CPN workshop (D1.1, 2.5) and encouraged by the industry experts at the WAN-IFRA session (D1.1, 2.4.2), we think this could increase the overall impact and the reach, beyond the publishers' traditional audience.

2.4.1. THE PROBLEM

Let's consider articles, videos and other news items as single units that make up a publisher's content offering.

That offering might or might not reach its intended audience, based on the publisher's choices and characteristics: the subject matter, the style of storytelling, the branding, the seriousness or sensationalism of the pieces and their complexity.

Current ways of personalisation exist only on this level, the level of the offering. They assemble a personalised offer using fairly basic criteria like:

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

- Subject matter of the news items
- the item’s metadata (tags, categories)
- Personal usage data
- Demographic data
- Usage data of peers
- Etc.

The news items themselves aren’t altered in any way, although they contain several useful elements and attributes. If we refine personalisation to that level and use more sophisticated criteria, we could allow for an even more targeted approach.

2.4.2. PERSONALISATION ON A DEEPER LEVEL

All news items consist of discrete elements and useful attributes. For the sake of this example we’ll use an article, but the same applies to news videos, audio snippets, and other news items.

The elements of a generic news article are:

- Title
- Header image
- Intro
- Body text
- Body images
- Related articles
- Author

Some of its attributes:

- Length
- Tone of voice
- Complexity
- Sentiment (rather positive or negative writing style)
- Required knowledge of the subject’s context
- Etc.

Some examples of how these elements could be adapted:

- Title: add two versions with a more serious and a more sensationalist tone, either manually or suggested by the system
- Header image: add several possible images to the same article. The simplest way to capture the difference between images is by manually or automatically tagging each image. E.g. whether a face is present in the image, using keywords (*car*, *child*, *nature*, *soccer*) or even names of celebrities). The system then has to predict the right image for each user. E.g. young parents might like images of children, teenagers like pop stars, … Netflix also uses this approach to personalise the artwork of their films and series[24].

[24] Artwork Personalization at Netflix: https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76

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

- Intro: add a second intro in which you highlight another part of the article. E.g. "This is the 20th school shooting this year" vs. "The shooter had mental health problems".
- Body images: display images in the article's body text to break up long paragraphs of text, or exclude them depending on what the user finds easier to read.
- Related articles: offer a choice between more of the same, more context, newer articles or very old ones, depending on a user's knowledge of the subject
- Body text: since this is the core of an article, adaptation is a sensitive issue and met with a lot of scepsis from journalists. However, personalized versions of an article may increase the number of reads. Some users prefer a short and to-the-point version, whereas others want an in-depth analysis of the subject, possibly with references to other news items. The same goes for sentiment and complexity. Which article version an individual user prefers, depends on personal interests and their current context.
- Author: users could have a preference for the writing style and subject choice of certain journalists. Machine learning techniques could automatically find a perfect match, using data from all users without human intervention.

2.4.3. MORE SOPHISTICATED CRITERIA FOR PERSONALISATION

The criteria we currently use to personalise a content offering are fairly basic and high level. We group news items together based on their subject matter, metadata, the users' usage and demographic data, etc.

If we used more refined criteria, individually per user on an almost psychological level, we could improve the offering. Possibly advancing to an *audience of one*.

Two of such advanced techniques are sentiment analysis and difficulty level analysis.

Sentiment analysis

In D1.1 we described the co-creation session with the media professionals at VRT. One of the results was a user story involving persona "Anne" (see D1.1, 2.3.2). The professionals agreed this story was the most interesting one to develop further.

In short, the user journey states that Anne has the VRT NWS app but she is not really interested in "hard" or political news. The app observes Anne's behavior and knows which articles she opens. Continuously, the app tracks data about the words in the titles and body texts (e.g. emotional words, positive or negative emotion, anger, sadness, etc.).
This way, the app discovers that Anne is triggered by positive emotional words *and* that these affective words are predictive for her clicking behaviour and reading time.
Subsequently, the app starts to give Anne more personalised content. Not based on broad topics or tags, but on emotional (and other psychological) parameters.

This user story serves as a starting point. In a first phase, we will make a dataset with roughly three categories of variables: behavioural data, personal data and textual analysis data.
The behavioural data could consist of clicking behaviour, reading time and other variables of interest. Personal data could consist of socio-demographics, but also "deeper" personal data such as political viewpoints, values and moods.
At last, and of most importance for this section, is the textual analysis data. For that, we will use a Natural Language Processing (NLP) tool such as the Linguistic Inquiry and Word Count (LIWC;

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

Pennebaker, Francis, Booth, (2001). This sentiment analysis tool captures conscious and unconscious psychological phenomena related to cognition, affect, and personal concerns. LIWC is used in a variety of disciplines and has proven to be a valid tool (Pennebaker et al., 2007). However, other sentiment analysis tools are also available and can be tested for our purposes.

A sentiment analysis tool enables us to discover deeper insights into the writing style of an article. More importantly, we can discover which textual variables such as authenticity, emotional tone and analytical thinking predict reading time or clicking behaviour, and for which type of user (for a full list of variables see Pennebaker, et al., 2015). So for our dependent or outcome variables (reading time or clicking behavior) we will conduct linear or logistics regression analysis with authenticity, emotional tone and analytical thinking, etc. as predictor variables and our personal data as moderator variables. This will enable us to detect important textual factors in an article that a certain reader (unconsciously) likes or doesn't like. In a second phase, analyses can be done on a personal level to discover what kind of textual factors predict reading behaviour for this specific news consumer.

This more profound way of analysing the textual characteristics will create another opportunity. While authors write articles, their text can be scanned automatically. Authors can receive immediate feedback on the emotional tone of their article and the algorithm might give an overview of the types of readers who will be interested (with some percentage points) in their article. The system might give suggestions to change the tone of the headline or add another version of the intro to reach a bigger audience.

This resembles the way modern advertising platforms like Facebook Advertising Manager work. They too try to predict how many people and what type of users you'll reach.

Difficulty level

Articles can be written in different styles, with different levels of difficulty, or accessibility. A short article with simple sentences and common words is easier to digest than a long, academic analysis with specialized jargon. Luckily, text complexity can be measured relatively easy. Metrics are for example:

- number of characters or words
- average length of sentences
- average length of words
- number of different words

Alternatively, text complexity can be computed with machine-learning techniques that compare the article with other texts that are labelled with a complexity score. However, this requires a substantial database of scored texts and more computational power, so that might be overkill.

The combination of the article's subject (tags) and the text complexity will probably yield the most successful prediction of whether a user likes an article. The reason is the individual background of a user in a specific field. For example, a scientist will desire an elaborate article on climate change but prefer a shorter one on the next elections.

We believe that offering several versions of the same article, with different levels of difficulty, in function of what we expect the user prefers, will greatly improve the reading experience.

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

3 CONCLUSION

The document has outlined possible innovative components that can be used in CPN’s technical framework.

It has provided the reasoning why these components were chosen and a detailed description of each. We believe these components were thoroughly rooted in the exploration phase, and came from our understanding of the users, both publishers and their audience

The document acknowledges the state of the art of current personalisation approaches, and explains why and how they would add value to CPN.

The consortium partners will discuss opportunities that arise from this document and will work together to validate them in the period leading up to the pilots.

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