Grant Agreement No.: 761488

D4.3: Cycle 2 Piloting Report

This deliverable contains the details of all activities around the preparation for the second pilot prototype and its validation.

Work package	WP 4
Tasks	T4.1, T4.3, T4.4
Due date	20/09/2019
Submission date	27/09/2019
Deliverable lead	Deutsche Welle
Version	1.0
Authors	Wagner, Tilman (DW); Kisselmann, Olga (DW); Mattheijssen, Joris (VRT); Van Lier, Jens (VRT); Coppens, Paulien (Imec); Danezis, Christos (Dias); Konstantinou, Stavria (Dias); Harrison, Jamie (Digital Catapult); Demeester, Thomas (Imec)
Reviewers	Lemmelijn, Ilke (VRT); Mattheijssen, Joris (VRT); Van den Broeck, Wendy (IMEC)
Keywords	Pilot 2, Evaluation, User Testing

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	24/06/2019	First draft version	Wagner, Tilman (DW)
V0.2	23/08/2019	Input added by partners	All partners
V0.3	30/08/2019	Review of Input and refinement	Kisselman, Olga, Wagner Tilman (DW)
V0.4	13/09/2019	Finalising for review (adding missing input)	Kisselman, Olga, Wagner Tilman (DW), all partners
V0.5	24/09/2019	Integration of Review Comments	Kisselman, Olga, Wagner Tilman (DW)
V1.0	27/09/2019	Final Version	Kisselman, Olga, Wagner Tilman (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.

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**

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

EXECUTIVE SUMMARY

This deliverable provides an overview of the different steps that have been undertaken to improve, test and validate the CPN concept, platform and mobile application after the completion of Pilot 1. The activities include Pilot 2, which was conducted by all three media partners, as well as supplementing evaluation, prototyping and research activities
This deliverable is the successor of D4.2.

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

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

LIST OF TABLES

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

ABBREVIATIONS

IP	Internet Protocol
TCP	Transmission Control Protocol
PoC	Proof of Concept
SSO	Single Sign On

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

1 INTRODUCTION

The focus of this deliverable is to report on the second CPN pilot. The pilot encompasses the development of the second version of the CPN prototype, including all experimental add-ons and further research on features as well as the evaluation of the final prototype in a large-scale living lab setup as described in D4.2 Cycle 1 piloting report.

Furthermore, this document encompasses all the activities that were done along the process of the development of the second prototype on the research side to further improve the application and test different approaches and ideas. This refers to the side-tracks that were designed and realized by partners VRT and DW, the exchange and cooperation with external parties (such as the PEACH project), as well as feature deep-dives performed by the media partners to further explore new ideas and concepts. More details will be given in the following description and the respective chapters.

The deliverable structure follows the idea of briefly explaining the connections with other (technical) developments and documents and the overall requirements scheme to clarify what requirements have been added in the second pilot (chapter 2). It then takes a closer look at the aforementioned focus research undertaken by media partners and explains the various set-ups, results and take-aways in more detail (chapter 3).

Following is a detailed description of the sidetracks that were done in the second year (chapter 4), covering the idea behind each side-track, the set-up, evaluation method and results. It is also explained how the results (as far as already available) affected the further progress of the Pilot 2 development.

All previous steps lead to the Pilot 2 evaluation description (chapter 5). First the set up and goal of the evaluation are explained, including a recap of the methodology (as already described in detail in D4.2), as well as the recruitment process of the participants. Then the results from the actual piloting are displayed and explained in detail. Starting with an in-depth analysis of the usage of the application during the testing period, followed by the direct feedback from the user surveys and the results from the focus groups. The chapter concludes with the main findings of the Pilot 2.

Finally, the process is further extended with the results from the hackathons (chapter 6). A short overview of the ideas and solutions proposed by the participating start-ups and selected for CPN is given, together with a description of how the events were organized and how the results will be integrated into the CPN process.

The deliverable closes with a short summary and an outlook on the next steps (chapter 7)

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

2 CPN APPLICATION DEVELOPMENT & REQUIREMENTS

The basis of a successful evaluation is the continuous development of the prototype and its features. This development happens in equal parts between the functional and the technical side. In order to make it clear which technical developments happened in preparation for the second pilot, the following chapter will give a brief look into these links, connecting this deliverable to the other relevant ones.

The second pilot execution is based on the open virtual platform v2, an advancement of the first version, released for pilot 1 in year 1.

In particular, the second version of the platform serves to:

- evolve the core components
- implement a new set of features through the deployment of new technology bricks and updates of the existing ones
- improve the reliability, security and performance of the overall architecture

The complete list of the activities conducted, the updates of the platform and the technology bricks deployed were reported in D2.3 - CPN Open Virtual Platform v2[1].

Below is reported the process of features implementation, is reported on, based on user requirements expected for the Pilot 2.

2.1 REQUIREMENTS COVERED

In order to satisfy the requirements expected for the second pilot (as described on D1.4 Technical requirements[2]) an agile Scrum-like methodology was applied.

At the start of the process, the user partners (VRT, DW, DIAS, IMEC), defined a prioritization for the requirements. In a second step, the technical partners mapped the technology bricks with the requirements, in order to identify the responsibilities of the technical activities. As a third step, an iterative activity (two-week sprints with planning and release at the end of each period) was followed for the implementation.

1 https://www.projectcpn.eu/s/CPN_D23_CPN_Open_Virtual_Platform_v2_20190530_v10.pdf

2 https://www.projectcpn.eu/s/CPN_D14_Technical_requirements_platform_and_service_requirements_20180830_v10.pdf

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

The result of this process was the delivery of a series of new technology bricks, the improvements of the existing ones and the deployment and the integration of all of them within the CPN platform.

Below is the list of requirements that were covered, discarded or postponed for further evaluations, as described in D2.3:

Table 1: Overview of Requirements for the Pilot 2 prototype

Req. ID	Description	Status
UR-UP2.1	The system should allow for social media integration to recommend content based on what connections like, read and share.	**Completed**
UR-UP2.2	The system should offer a recommendation of articles based on most liked/most shared numbers from a user's network and beyond that. (Nuzzle-Feature)	**Completed**
UR-UP2.3	The system should allow for social media integration to keep track of what the user has already seen elsewhere	**Completed**
UR-UP 2.4	The system should be able to analyze whom a user has been most interacting with on social media to prioritize the users for the personalization on social media to prioritize the users for the personalization.	**Completed**
UR-UP 2.5	The system should allow the user to down-/ upload their network connections through user account.	**Completed**
UR-UP3.1	The system must allow the user to choose a preferred time frame or frames to consume content	**Completed**
UR-UP3.2	The system should create/refine time frames based on the user's consumption habits	**Completed**
UR-UP3.3	The system should refine the user's time frames through frequent interaction with the user (talkback)	**In evaluation**

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

Req. ID	Description	Status
UR-UP3.4	The system should use the time frames in order to decide how many items of what length and of what format it offers to the user	**Completed**
UR-UP3.7	The system should learn from these user responses and adjust its offerings accordingly	**In evaluation**
UR-UP5.1	The system should make use of the location data of the user (permission of the user granted) to choose the right content for the user	**Completed**
UR-UP5.2	The system should allow the user to set a home/ main interest location.	**In development**
UR-UP5.5	The system must give the user an easy option to agree to or withdraw from using location data for personalised offers	**Completed**
UR-UP6.1	The system must keep track of what content the user has already consumed on a piece and on a content basis within CPN and beyond	**Completed**
UR-UP6.2	The system must keep track of how much of each item users consume, where they stop, continue and what they skip	**Completed**
UR-UP6.3	The system should interact with the user in order to refine user interests in regards to why something was skipped or something was consumed completely	**Completed**
UR-UP8.2	The system should always offer content that has a direct influence on the users (e.g. life-threatening), overruling other interest settings	**Completed**
UR-UP9.3	The system must give the user a full overview of his/her data and allow them full control, including update and removal of data	**Completed**

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

Req. ID	Description	Status
UR-UP9.4	The user must be able to change and overwrite settings in their profile	**Completed**
UR-UP9.5	The user must be able to download their profile data in CPN in a machine-readable format and a user-friendly format	**Completed**
UR-AF1.6	The system should offer the user a random news selection upon request based on certain data and preferences of the user's profile, which the user can choose	**Completed**
UR- AF2.5	Once all articles proposed have been consumed, the system should only offer more content upon request by the users	**Completed**
UR-AF5.1	The system must offer the user an easy access and easy to understand overview of their profile	**Completed**
UR-AF5.2	The system must offer users easy access to their profile in order to change settings and data	**Completed**
UR-AF5.3	The system must make it transparent to the users why they are shown certain content, based on an item level	**Completed**
UR- AF6.2	The system should allow users to consume content beyond their predefined time- frame after an interaction with the user (talkback)	**In evaluation**
UR- AF6.3	The system should allow users to actively save articles for later consumption	**Completed**
UR- AF7.2	The system should include guided feedback for specific elements of the system, allowing users to (help) improve it	**Completed**

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

Req. ID	Description	Status
UR-AF8.1	The system should allow users to search for specific topics they are temporarily interested in	**Completed**
UR-AF8.2	The system should allow users to add this search as a temporary personalisation category	**Completed**
UR-AF8.3	The system should allow users to define a specific time frame for this temporary change	**Completed**
UR-AF9.1	The system should allow users to define keywords and logical combinations of them to exclude content from their personalisation	**Completed**
UR-AF9.2	The system should allow users to define a time frame per keyword/logical combination	**Completed**
UR-AF9.3	The system should be able to overwrite this exclusion for important breaking rules	**Completed**
UR-PS1.1	The system should show the access to items through users by numbers (who, when, how long)	**Completed**
UR-PS1.3	The system should show which topics were most interesting to users	**Completed**
UR- PS2.4	The system should allow producers to export the record of their publications through standardized and interoperable formats	**Completed**
UR- PS2.5	The system should allow for an easy contribution of content from different publishers through standardised interfaces	**Discarded**
UR- PS2.7	The system should allow editors to easily add missing attributes to articles manually	**Discarded**

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

Requirements in development

The only task still in development phase is the UR- UP5.2. In particular, the user can save/update a preferred location through the UI of the client application and this information is saved and available for the CPN platform components.

At the moment, this information is not yet used by the recommendation engine because it requires the content analysis for the identification of position information from the article. This functionality is not yet implemented and will be added for the third version of the platform.

Depending on priorisation, the next version of the CPN platform if foreseen allow the user to choose different kinds of location:

1. Main location
2. Other locations (to be added manually)
3. Location tracking (with user permission)

Requirements under evaluation

The requirements under evaluation and postponed to the third version of the CPN platform are: UR-UP3.3, UR-UP3.7 and UR-AF6.2.

All these requirements concern the "Time Frame" macro-functionality which, in a basic version already implemented, allows the users to consume content in a given period.

In order to satisfy the aforementioned requirement an evolution of the functionality is expected but under evaluation.

According to early feedback collected from users, this feature has been little used and has not had much impact, moreover the effort required to implement an advanced version is very high.

Discarded Requirements

At the end of the second CPN platform implementation, two requirements were discarded: UR-PS2.5 and UR-PS2.7.

Both of these requirements are linked to in the editing of the articles through the UI of the Producer's App (the Producer's Dashboard).

Both the internal media partners of the consortium (VRT, DIAS, DW) and the media companies so far involved in the integration of their systems with the CPN platform, categorically excluded the possibility to edit articles on a platform other than their CMS.

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

In this case, the integration with media companies will be only in a as-a-service approach (as described on D2.3 CPN Open Virtual Platform v2[3]) and the as Producer's Dashboard will offer analytics and data on user consumption.

3 https://www.projectcpn.eu/s/CPN_D23_CPN_Open_Virtual_Platform_v2_20190530_v10.pdf

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

3 FEATURE DEEPDIVES

During the requirements process at the beginning of the project, the consortium identified some basic features necessary for a successful personalisation algorithm. While this set of features worked fine for the creation of the first CPN prototype, the evaluation made it clear that some features were more complex and needed more research.

As a result, the consortium came up with a plan to further explore the options around these features and also to test various alternatives to the existing system setup.

The following chapter describes both the initial plan as well as the measures that were undertaken in year 2. The plan for the so called "Feature Deep-Dives" (formerly known as "Mini Pilots") was set up as an analysis stream to cover several topics and concepts in parallel to the ongoing development of the second CPN prototype (app stream and service stream). Split across the pilot phases, it comprised of 8 separate modules for the 2nd and 3rd year (Figure 1)

CPN	Pilot 1	Pilot 2	Pilot 3
App Stream	News App v1	News App v2	News App v3
	Clickable Wireframes	/ (integrated into app)	/ (integrated into app)
		Personalised Newsletter v1	*Personalised Newsletter v2*
		Web Integration v1	*Web Integration v2*
			News Bot
			Smart Speaker
Services Stream	Recommender v1	Recommender v2	Recommender v3
	Personal Data Receipts v1	Personal Data Receipts v2	Personal Data Receipts v3
	DLT Licensing v1	DLT Licensing v2	DLT Licensing v3
	Rest API v1	Rest API v2	Rest API v3
		Analytics on News Items v1	Analycits on News Items v2
		Sentiment Extractor v1	Sentiment Extractor v2
Analysis Stream	General Acceptance	Alternative personalisation strategies	Including Seredipity
	Feedback Questions	Mapping Strategies (keywords vs. interests)	Best Practices for Personalisation
	Cold Start vs Categories	How to handle Breaking News	Editorial prerequisites for good personalisation
	News Summaries vs Background Info	Breaking the Filter Bubble	
	Location based news	Fighting FOMO	

Figure 1: Overview of Feature Deep Dives as planned during year 1, next to the overall development

The idea was to take on those topics one by one and explore the options through in depth research, small experiments, lab stages and user-testing wherever possible. The results were supposed to be evaluated by both technical and user partners to make a decision on which ones could be integrated into the main application development to further improve it.

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

For year 2, it was foreseen to take a closer look at the following topics:

- Breaking News
- The filter bubble
- Different mapping strategies
- Alternative personalisation strategies
- The fear of missing out (FOMO)

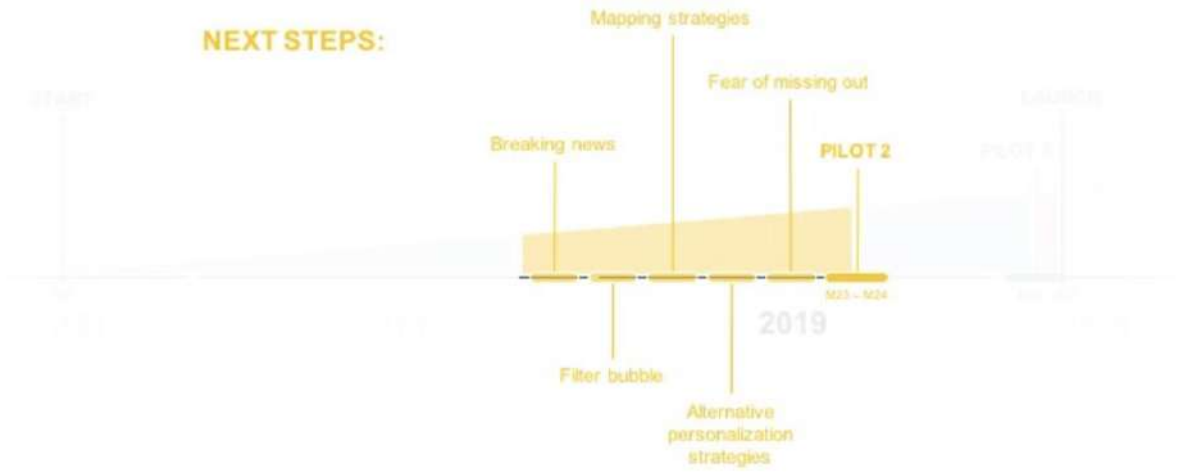


Figure 2: CPN Feature Deep-Dive Timeline for year 2

In the initial plan, a slot of two months was envisioned for each topic, allowing enough time to draw conclusions from the different stages and possibly integrate results in the main development. While this plan worked well for some topics, in other cases it didn't work and the initial time-plan was handled in a more flexible way. Some of the topics found their way into the hackathons and were taken on by the participating start-ups. Others proved too complex to be conclusively solved in this short amount of time. Overall the consortium was able to learn from these deep dives, which it also discussed with external partners and project stockholders at the different events undertaken in year 2.

The following subsections give a detailed overview over the constituent research blocks, the outcomes or reasons for postponing/moving them and how they were used further in the process.

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

3.1 BREAKING NEWS & PUSH NOTIFICATIONS

As per the initial definition, 'breaking news' is "information that is being received and broadcast about an event that has just happened or just begun."4 In the age of television, breaking news would lead to an interruption of the linear program to inform the audience about something new and important. Once the news was out, people would go back to the normal program. Since it happened seldomly and through people's main (and sometimes only) news channel, everyone was interested in it.

In the age of 24-hour news-cycles, the label has become somewhat overused and irrelevant, as almost everything is new and breaking with the daily routine. In addition, media companies are in the middle of a harsh fight over our attention with social media and other "news" providers. This has led to an annoyance of users, especially since smartphones have become people's main source of information. Push-notifications, often the audiovisual messenger for breaking news, have taken over the screens and have added to people ignoring them, suppressing them or, in the worst case, deleting the dispatching app.

From a personalisation perspective, it quickly becomes clear that "breaking" is simply not the same for everyone. If news is personalised, why should breaking news and accompanying push notifications be treated any different? Taking a closer look at the question however reveals that there is no straightforward solution to this task. There are several core questions that need to be answered for a good solution:

- What makes news breaking?
- Are there breaking news items that are so important to everyone that they need to completely overwrite the personalisation settings?
- If not, what are the rules to overwrite or to be overruled by the algorithm?
- Is there a middle way to not overwrite the personalisation and still make it clear to users why they are receiving certain news without them losing trust in the algorithm?

In order to get a better understanding of the situation and to be able to find answers to those questions, the CPN consortium decided to devote one deep-dive set to breaking news and push notifications. In cooperation with Wan-Ifra, DW launched a short survey among media professionals and newsrooms for a better understanding of the journalists view of breaking news. In addition, DW took a closer look at its own breaking news and push-notification concept in a two weeks DW Lab session.

4 https://dictionary.cambridge.org/de/worterbuch/englisch/breaking-news

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

3.1.1 Survey among Newsrooms

The survey was set up as a short questionnaire with two parts, one on breaking news, the other one on the use of push-notifications. The questions were defined in collaboration between DW and Wan-Ifra, focusing on the experience of newsroom journalists and their daily routines in dealing with breaking news.

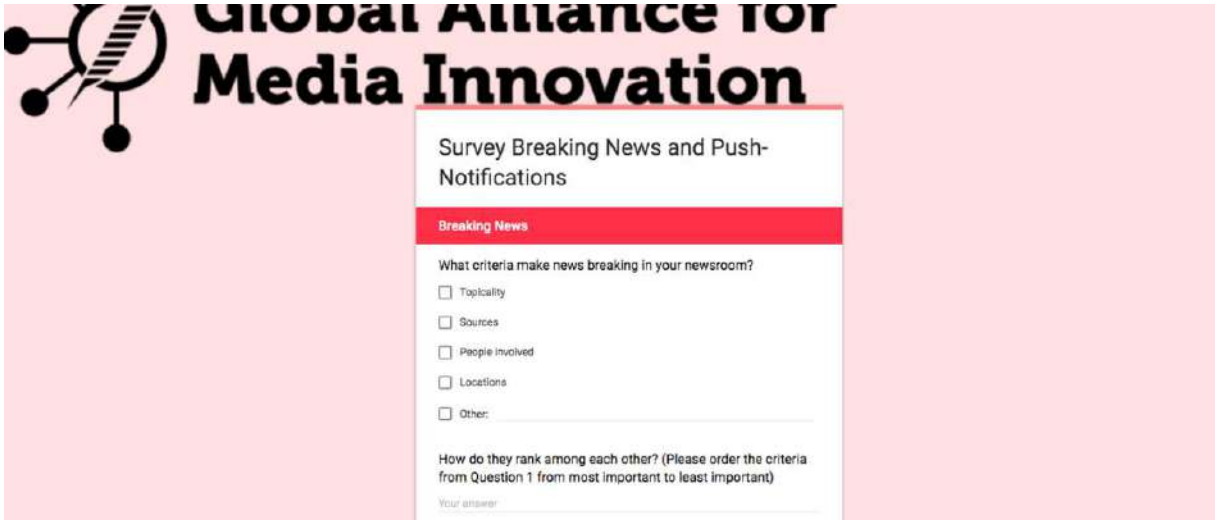


Figure 3: Example from the survey sent out to media companies

For breaking news, there was a total of eight questions, plus a field for comments. The questions focused on the decision-making process and workflow dealing with breaking news within newsrooms in comparison to regular news. It included questions regarding the roles involved and the level of standardisation/automation to better understand how a personalisation approach would have to be designed to cover all important aspects and deliver high quality results.

The set of questions regarding push-notifications encompassed a total of ten questions plus a field for comments. The focus of this section was to analyse the connection and the strength of the link between breaking news and push-notifications. Again, it included roles and responsibilities as well as automation and standardisation.

With the final version of the questionnaire, Wan-Ifra reached out to its network of media companies, asking for volunteers to participate in the survey. Over a period of two weeks, eight different individual responses could be collected (see Figure 4 - one newsroom didn't want to be named) from a diverse set of media companies from across the globe.

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

Affiliation/organisation

7 responses

Dagens Næringsliv	Norway
Süddeutsche Zeitung	Germany
TheStreet	USA
INM	Ireland
Imagen Laguna	Mexico
TheQuint.com	India
BBC World Service	UK

Figure 4: List of organisations participating in the breaking news survey (except one)

3.1.2 DW Lab Stage

The DW Lab is a safe space inside Deutsche Welle's innovation department for technical and user-centered media experiments. It is used to develop new formats, test new technologies and analyse existing methods and workflows. Members of DW can apply for a two-week slot to work on a specific topic, in-line with DW's digital strategy. The sessions are moderated by a dedicated team and in collaboration with other departments, involved in the topic at hand.

The DW-CPN team applied for such a slot to take a closer look at DW's current handling of breaking news and push notifications and got the go ahead from the Lab. In preparation for the session, the team took a closer look at state-of-the-art methods. During the two weeks' timeframe, the DW CPN-team further analyzed different media companies' strategies, took an analytic approach to its own push-notifications and developed and tested a prototype to gather further user feedback on new approaches.

For the state of the art, the team did an in-depth research into technical literature and web articles. The majority of the found posts and articles found were more concerning the advertisement market (e.g. push notifications for booking- or shopping apps), but they already gave an indication to what was currently "en vogue": Not too much, not too often and better personalized than "one fits all". The analysis of media competitors (via screen-capturing other media outlets' push notifications) gave a good insight view into aspects that were already clear from the breaking news survey: There is no clear logical rule for what makes breaking news. But it also revealed that the trend among media companies goes already more in the direction of allowing the user a more personalized news digest, e.g. by letting them choose topics of interest for which they want to get more info. Still it all seemed to be experimental on basic levels only.

The analysis of DW's own breaking news and push notifications can't be further disclosed herein, but the insights led to the design of a prototype valuing the users

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

wishes for more personalisation possibilities in regards to when, how long and about what they would be informed. The prototype was tested at DW internally and received good feedback regarding the concept and setup.

3.1.3 Breaking News Summary

The survey among media professionals clearly showed that breaking news can't be handled strictly on a logical basis. None of the responding newsrooms had any formalised process to decide on what was breaking news. The decisions relied heavily on the experience of the staff and the current news situation with topicality and people involved (in the news) as the top two criteria to decide upon, followed by the respective source, the news were coming from. This of course made it difficult to come to a clear solution on how to best set up a personalisation logic in regards with handling breaking news and respective push notifications.

What was clear through the research however: breaking news and push notifications are better received in a personal way. How exactly this should happen depends on the newsroom's overall strategy. Following the discussions with media companies and looking at the results of the survey, newsrooms don't like the idea of overwriting their own decisions. If they decide on something being breaking, they want people to see it. They set the bar for this very high for themselves, which is why they think it shouldn't be overwritten.

Trying to bring both of these aspects together is the middle ground that CPN needs to find. A starting point is the indication of priority inside a personalisation app: If breaking news should overwrite a personalisation algorithm, then only if the topic is important to the user, or the people involved as a second, or the source or location as a third and fourth characteristic.

Taking the results and learnings from this session into consideration, DW created a flowchart for a possible overwrite logic (see Figure 5). It assumes that all necessary indicators for the personalisation can be extracted or measured without errors. It is also based on the assumption that there is a hierarchy in what elements are more important than others, overwriting the others. When something is of utmost importance for the user (e.g. in immediate vicinity, life threatening etc., then it's clearly overwriting the personalisation. If it doesn't fit into user's profile, following the order taken from the survey, then it should not be shown.

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

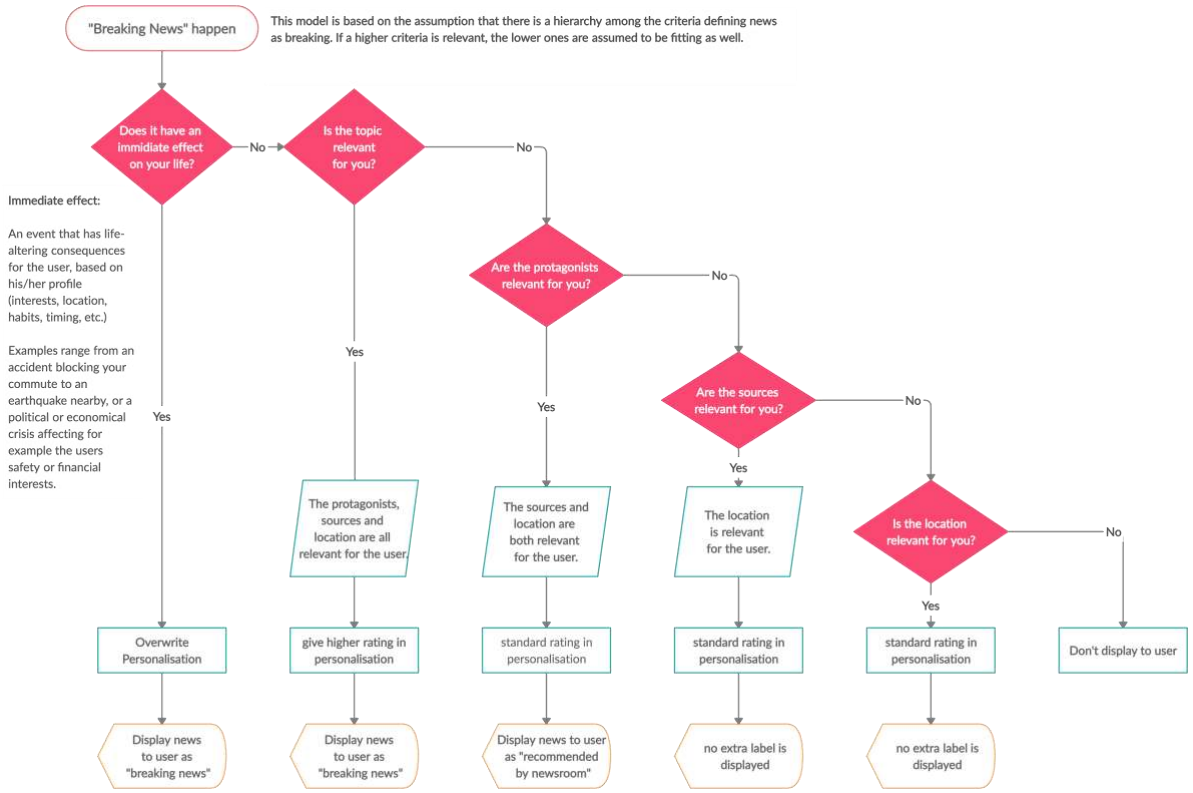


Figure 5: Decision tree for handling breaking news in a personalised news offer

Right now, CPN doesn't have extra push-notifications for breaking news, as the consortium values the users own time frame settings as more important. There is already a discussion ongoing in the consortium to work with extra labels (as described in the Figure above); a definite solution will have to be addressed in the third version of the prototype and evaluated in pilot 3.

3.2 BREAKING THE FILTER BUBBLE

Since the beginning of the project, we have discussed and researched a possible negative impact of news personalisation. An often-quoted concerns are the so-called informational filter bubbles and echo chambers, that may occur as a consequence of extensive algorithmic personalisation. The concept of filter bubbles was already known in the early 2000s but was widely popularised by the internet activist Eli Pariser in his 2011 publication " The Filter Bubble: What The Internet Is Hiding From You"[5].

[5] https://books.google.de/books?hl=de&lr=&id=-FWO0puw3nYC&oi=fnd&pg=PT3&dq=The+Filter+Bubble:+What+The+Internet+Is+Hiding+From+You%E2%80%9D#v=onepage&q&f=false

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

Throughout the project, we have analysed various aspects and possible solutions to the filter bubble problematic with the intention to:

1) Asses the potential risks of algorithmic personalisation.

2) Create monitoring and mitigation measures to ensure that our application will perform above existing standards, increase overall media literacy of our users and exposure to diverse content while still providing a personalised experience.

3.2.1 Filter Bubble and Echo Chamber Theory

According to Pariser[6] and other proponents of this theory[7], algorithmic filtering of news content may create informational loops where only favorable content, which is consistent with existing opinions and believes reaches the user and hence increases the *confirmation bias*[8], while filtering out perspective-challenging information. As a consequence, this can increase political and social polarization. While the Filter Bubble theory has gained significant media attention and is today still strongly associated with the rise of populist movements and politicians around the globe, but especially in Western societies with stable democratic traditions[9], the impact of algorithmic filtering on political polarisation is still open to debate.

To our knowledge, there is a number of research contributions that would strengthen the case against a significant impact of personalization algorithms on both, overall news diversity exposure and political opinion-making. Those include observations from the social media networks[10], standalone multi-source news applications[11] and the usage of different algorithmic recommendation engines, including collaborative, topic-based and individual behaviour-based content recommendations on single-source news content. We consider Möller et al. (2018)[12] to be particularly important for our project, because of the similarity of the research design to our use case.

6 https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles?language=de

7 https://www.theguardian.com/technology/2017/may/22/social-media-election-facebook-filter-bubbles

8 https://bigthink.com/Charles-Koch-Foundation/facebook-algoithm-filter-bubble

9 https://www.wired.com/2016/11/filter-bubble-destroying-democracy/

10 https://www.uni-hildesheim.de/wahlkampfanalyse/; https://gfx.sueddeutsche.de/apps/e502288/www/; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6214567/; https://www.niemanlab.org/2017/06/using-social-media-appears-to-diversify-your-news-diet-not-narrow-it/; https://cristianvaccari.com/2018/02/13/how-prevalent-are-filter-bubbles-and-echo-chambers-on-social-media-not-as-much-as-president-obama-thinks/

11 https://journals.sagepub.com/doi/abs/10.1177/1077699013514411 /10.1287/mnsc.2013.1808 https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2013.1808

12 https://www.tandfonline.com/doi/full/10.1080/1369118X.2018.1444076

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

3.2.2 Creating content-diversity-oriented features for the CPN Platform

At the current stage of our research, we came to the conclusion that when applied to our use-case, the risk of creating content bubbles -that would strongly limit the diversity of perspectives, cut out important information and run the risk of increasing political and social fragmentation through personalisation technology- is relatively low. While assessing the potential risks, we have taken into account our primary use-case, which is a customizable personalization application for individual media houses. At this stage of the project, we are only dealing with a single-source content scenario. This would change, if more publishers than the current content partners would use the platform and opt for a multi-source content CPN application. Within the single-source scenario, we don't see the risk of political polarisation that would go beyond the initial limitations of the editorial perspective of any given media house. Instead, we believe that consistent with Möller et al. (2018), users might over time be exposed to more topics and content items.

However, we understand that the given issue is not sufficiently researched to dismiss any potential risks of algorithmic filtering in the future. At the other hand we also aim to create additional value by increasing media diversity, which is the overall goal of CPN. Therefore, we have conceptualized and implemented several functionalities in the CPN platform and the mobile application that will ensure that users will be informed about news of significant collective importance. For this reason, the CPN mobile app has three different content streams, of which only one has individual content recommendations in place. The other two streams will ensure that the content of significant collective importance will still be visible for all users. In section 3.2 that presents our research on breaking news and push notifications, we have been analyzing different ways to optimize and strengthen (non-algorithmic) editorial control. We have also been experimenting with different labels for news content items that need an increased visibility, like breaking news. However, at this stage of the project, we didn't prioritize the implementation of those labels as we find that the three different content streams are sufficient to showcase news of significant collective importance.

With the increased transparency features that are described in section 3.9, we will ensure that our users understand the concept of personalized news and why certain content items are recommended to them. Additionally, the CPN app allows the user to see the content items they have read in the past. This feature is a crucial part of the application. As we are aiming for a transparency that sheds light on the whole process of news consumption, we want to encourage the users to take a look at the content they are consuming.

We are still working on a coherent approach to increase the user's agency. For Pilot 3 we envision either a user-facing visualization that would show topics and categories the users are consuming and possibly those that are underrepresented or an added instruction that would encourage the users to take a look at their own news consumption on a regular basis.

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

3.3 MAPPING STRATEGIES (KEYWORDS VS. INTERESTS)

At the core of every personalisation algorithm is the mapping of media items to a user. The better this matching works, the happier the user is (in general). While in theory this "just" means to find out what a user likes to read/watch/hear, it's a bit more difficult to solve on a technical level, both in finding the proper attributes in user behaviour and media items as well as matching them properly.

Since there are many different possible approaches to this question, the consortium wanted to take a closer look at some of the options. In particular the aim was to see what possibilities there are for improving recommendation quality by standardizing categories and keywords. The definition of both categories and keywords is not a trivial thing and media companies are struggling with this issue especially now, when everyone is trying to automate a lot of the processing of media items for better curation and/or recommendation on websites or in apps. Even with some algorithmic support, setting keywords is always a somewhat faulty process and prone to errors. It's a problem DW is working on for a while now.

During the deep focus on this topic, the user partners took a closer look at existing structures of other media companies, comparing categories for news (Figure 6). It quickly became clear that there is no unified categorisation in place. This poses a problem when it comes to building one system across media companies, especially when relying on media categories as a possible mapping criterion. The way around this is to create an internal mapping of external to internal categories. Further research showed that this is not an easy task, which was also confirmed through DW internal expert interviews. A fixed set of categories doesn't give the flexibility the news environment seems to need, as there is always something new coming up.

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

Deutsche Welle

1. Germany
2. Brexit
3. World
   - 3.1. Europe
     - 3.1.1. Germany and Turkey
   - 3.2. Africa
     - 3.2.1. Crime Fighters
     - 3.2.2. The 77 Percent
     - 3.2.3. Africa on the Move
   - 3.3. Asia
   - 3.4. Americas
   - 3.5. Middle East
4. Business
   - 4.1. Founders Valley
   - 4.2. Wo+men
   - 4.3. My 2030
5. Science
6. Environment
   - 6.1. Global Ideas
     - 6.1.1. In focus
     - 6.1.2. DoingYourBit
   - 6.2. Eco@Africa
   - 6.3. Living Planet

VRT

1. Rubrieken
   - 1.1. Binnenland
   - 1.2. Buitenland
   - 1.3. Cultuur & Media
   - 1.4. Economie
   - 1.5. Ook dat nog
   - 1.6. Analyse
   - 1.7. Opinie£
   - 1.8. Politiek
   - 1.9. Justitie
   - 1.10. Wetenschap
   - 1.11. Podcast
   - 1.12. Sporza
2. Regionaal
   - 2.1. Antwerpen
   - 2.2. Limburg
   - 2.3. Oost-Vlaanderen
   - 2.4. Vlaams-Brabant & Brussel
   - 2.5. West-Vlaanderen
3. Dossiers
   - 3.1. Brexit
   - 3.2. Jobs bedreigd bij Proximus
   - 3.3. Van oud naar nieuw
   - 3.4. Michel II

The BBC

1. Home
2. News
   - 2.1. Home
   - 2.2. Video
   - 2.3. World
     - 2.3.1. World
     - 2.3.2. Africa
     - 2.3.3. Asia
     - 2.3.4. Australia
     - 2.3.5. Europe
     - 2.3.6. Latin America
     - 2.3.7. Middle East
     - 2.3.8. US & Canada
   - 2.4. UK
     - 2.4.1. UK
     - 2.4.2. England
     - 2.4.3. N. Ireland
     - 2.4.4. Scotland
     - 2.4.5. Wales
     - 2.4.6. Politics
   - 2.5. Business
     - 2.5.1. Business
     - 2.5.2. Market Data
     - 2.5.3. Global Trade
     - 2.5.4. Companies
     - 2.5.5. Entrepreneurship
     - 2.5.6. Technology of Business
     - 2.5.7. Connected World
     - 2.5.8. Global Education
     - 2.5.9. Economy
   - 2.6. Tech
   - 2.7. Science & Environment

Figure 6: Comparing different media categorisation systems (excerpt) for overlaps & differences

Since the CPN recommender is already set up to extract its own keyword from the articles it receives, and since the keyword-system at Deutsche Welle was not available for further analysis, the decision was made to further analyse the automatically extracted entities. LIVETECH provided the user partners with a raw sample set of articles and the related extracted entities (see Figure 7). DW took the data and analysed it in regards to complexity and quality.

http://www.sigmalive.com/news/local/523430	DIAS	pafu, kype, spyros georgiou, k georgiou, pafu, geo	epesemane de, o k
http://www.sigmalive.com/news/international/540616	DIAS	indonesia, kapota, wakatompi, sulauezi, tailandi, ki	ton iunio, e kina, e indonesia, mechri to
http://www.sigmalive.com/news/local/542129	DIAS	kuprou, kuprou ant, kupro, kupro, kuprou, kupro,	kuprou, cyprus, kuprou
http://www.sigmalive.com/sports/football/cyprus/futsal/53	DIAS	andora, kop, ael, omonoia, ael, omonoia, anorthos	amek kapsalu
http://www.sigmalive.com/news/local/542748	DIAS	lemeso, deuteras, astunomia	to proi
china-records-slowest-growth-rate-in-nearly-30-years/a-	DW	United States, Beijing, China	trade war, growth rate
germany's-alexander-zverev-exits-australian-open-/a-47	DW	Alexander Zverev, Milos Raonic	seed men's tennis player, fourth round, Canadian opponent
china-records-slowest-growth-rate-in-nearly-30-years/a-	DW	United States, Beijing, China	trade war, growth rate
10-un-peacekeepers-killed-in-mali-attack/a-47160849	DW	UN, Mali, Nusrat al, al, al Qaida	
israel-admits-strikes-against-iranian-quds-in-syria/a-471	DW	Syria, Israel, Iranian Quds, Damascus	rare acknowledgement, military action, Syrian state
german-mayor-slams-all-male-charity-event-for-refusing	DW	Bremen's, wouldn't	charity event, substitute mayor, men's club
colombia:-thousands-march-to-protest-cadet-bombings/	DW	Colombia, Duque	
oxfam-releases-global-inequality-report-amid-ongoing-c	DW	World Economic, Davos, Oxfam	rich meet, global inequality, controversial study
switzerland-remembers-protestant-reformation-leader-u	DW	Ulrich Zwingli, Zurich, Switzerland, Zwingli, Germa	ecumenical service
bundesliga-bulletin:-borussia-dortmund-and-bayern-mur	DW	Bundesliga	relegation picture

Figure 7: Raw excerpt of the original data set for extracted entity analysis

The analysis showed that the keywords can be categorised into four main areas: people, places, organisations and events. While most of them seem to be clear, the analysis also showed that the system doesn't always get it right. There is always a number of entities, that is either faulty, wrong or missing. (see Figure 8)

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

Figure 8: Example of extracted entities and their classification

A manual comparison also shows that the system doesn't always get the complete picture, missing out on some entities that would have been relevant for the right categorisation of the text or a better matching, as can be seen in Figure 9, added and highlighted in grey.

Figure 9: Example of missing entities in the list of articles

This result underlined the problem currently existing at media companies across the globe: Even with technical support, the output of entities (keywords, tags, categories) is a difficult issue and as for now not solvable perfectly without a combined approach of manual annotation and semi-automatic support.

As this topic proved to be more complex than what could fit into a deep focus session, the consortium had to think of other ways to further work on the topic. Luckily some of the startups participating in the hackathon had a focus on both trying to solve the issue of the unharmonised category catalogues in media as well as the issue of automating the keyword-process. Two of them presented ideas on how to address the topics and will be working on their solutions for the third-year pilot.

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

3.4 ALTERNATIVE PERSONALISATION STRATEGIES

The idea behind this feature deep dive was to take a look away from the mere interest-based content recommendation and evaluate other approaches. However due to other development and research tasks, the consortium decided against allocating time and resources on this in year 2.

Year 3 however will bring some opportunities to take a closer look at what's possible. This will be realised through the work done by the startups, building on their work in the hackathons as well as the modules developed by IMEC, namely the uplifting/depressing article classifier as well as the frame-based annotation module (used for location-based recommendation).

3.5 FIGHTING FEAR OF MISSING OUT (FOMO)

Fear of missing out is a phenomenon that has evolved from the overflow of information available to us all the time. Through the internet we are now able to access, consume and share vast amounts of information from all over the world. The "have you already heard the news" from former incidental neighborhood meetings has been overhauled by everyone posting and communicating on social networks and messengers all the time.

While this has made people more informed, it has also led to the feeling of not being able to stop. Endless streams of information, promoted and made popular by networks like Facebook and Twitter, have given people the anxiety of having to go on and on - or fear of missing out of something.

Research done by the consortium into this phenomenon has shown that there is a movement to change this habit under the slogan "time well spent". Apple for example has introduced a feature called "screen time" to keep an eye on how much time a user spends with their phone through a week, broken down to groups and even single apps. User can set timeframes when they want the phone to (soft) lock specific apps to reduce usage. Instagram has started to stop offering an endless stream of new items in a user's timeline, after they have scrolled through the list of all items fitting certain criteria. The total number is dependent on the number of users someone follows, the number of posts they have made within the last three days and how well these items fit the user's interests, according to Instagram's algorithms. Additionally, Instagram allows the users to set a reminder for how much time they have already spent in the app.

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

Figure 10: Screenshot of Apple's Screen Time Feature

None of these measures are forcing the user to quit using an app or device, but they are helping the users to be more aware of the time they have spent already on their devices. This is a path that CPN has also tried to follow, when first drafting its application in some more and some less sophisticated ways. In the wireframes, a "all caught up" message was envisioned. This screen was intended for users reaching "the end" of their recommended article list based on the timeframe they have set for themselves and an estimated reading time per article. This feature has not been fully realised and hence couldn't be evaluated yet.

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

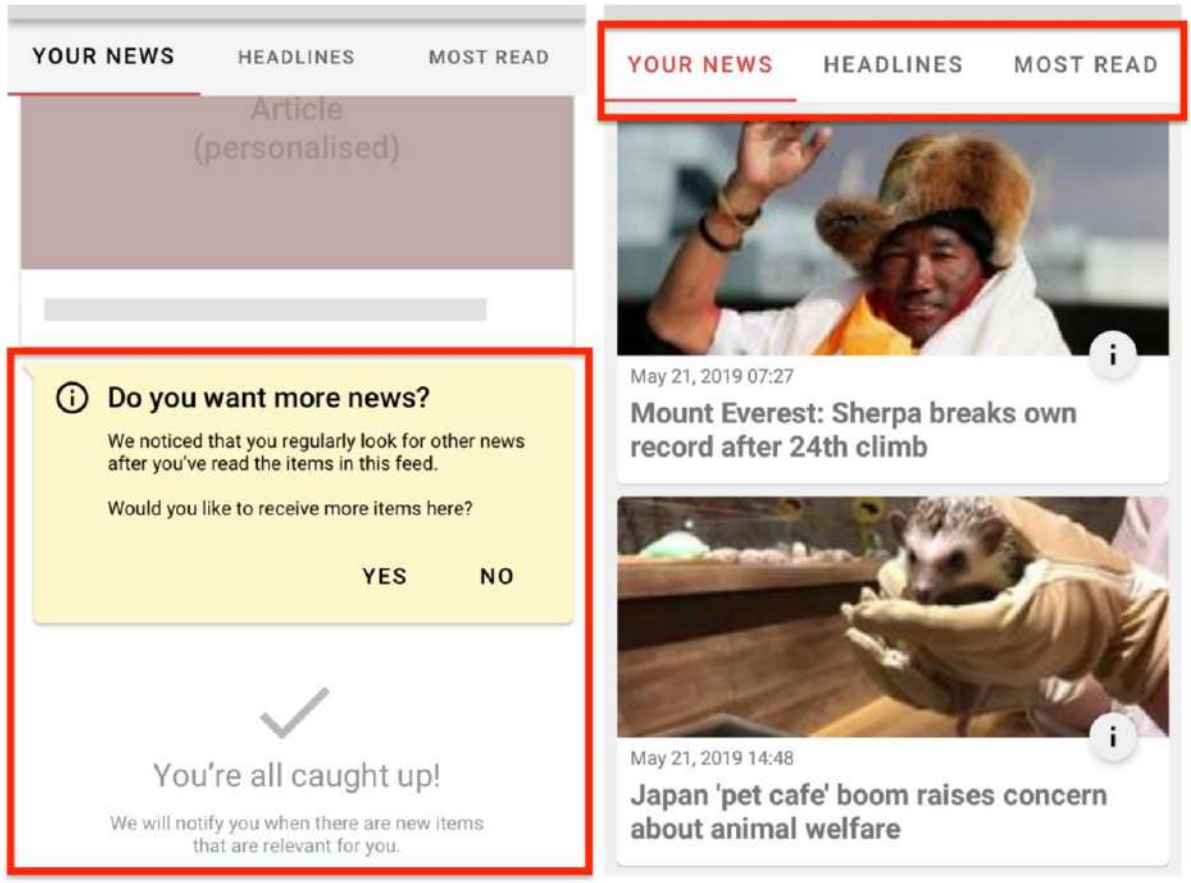


Figure 11: Examples of CPNs "time well spent" (wireframes, left) and transparency set up (three streams in mobile app, right)

Another, simple feature was the set-up of the app with three news-streams, one for personalised news, one for current headlines and one for most read news. The users are always able to switch between the three streams, allowing them to check, whether they are missing out on any topic. The idea here was to ease people's minds about using a personalised stream as their main source of information. Ideally, with a perfect personalisation, people would choose the stream that best serves their needs, ignoring the other two streams over time. As this feature had been realized in the app already, it was already part of the Pilot 2 evaluations.

The project did not set up another in-depth analysis or further evaluation of this topic, as the focus was shifted to other features due to conflicting resources. The consortium is currently discussing whether a separate evaluation can be arranged within Pilot 3 on these specific features and the effect on the user's behaviour. This requires of course that the preferred time frame and reading time feature is properly implemented.

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

3.6 LOCATION BASED NEWS

Even though the world has come closer together over the last decades due to developments in communication (and other) technologies, it is still the events happening nearby, that affect us most in our daily lives. This is why local news still holds an enormous relevance for people, even though headlines (and media) have become more global. In order to better understand how this reflects in users' preferences, CPN wanted to take a closer look at what role location plays in personalisation.

DW designed a first concept of location based personalisation, analysing the relation between a user's location (current or over time) and the news the user should/would want to receive. The difficulty here was to clearly define what a user means when they say that they are interested in a location.

There are several possibilities:

- News from media outlets in a specific location
- News about a specific location
- Mentioned directly
- Mentioned indirectly

The following concept was set up on a theoretical level and discussed among the user partners:

The system allows the user to set up 2-3 combinations of locations, they are interested in, in addition to allowing the app to track the current location.

The locations the user can choose from are based on four levels:

1. City (by name)
2. State (by name)
3. Country (by name)
4. Continent/Region (e.g. EU, Middle-East, …)

These choices get matched to the locations connected to the articles. Again, there are several levels that have to be taken into account:

1. City
2. State
3. Country

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

4. Continent/Region
5. Locations can be mentioned indirectly in the article
    a. Through a person (e.g. Angela Merkel, as Chancellor of Germany)
    b. Through an event (e.g. a festival taking place in a specific location)
    c. Through a part of a city (neighborhood, places, streets)/region with its own name
    d. The location of the journalist/media company (maybe not useful)
    e. Locations can be mentioned directly in the article as

The difficulty lies in finding these locations and then matching them to each other. The concept that seems to make most sense for this is the following:

Rules (what matches what)

		User Choice			
		City	State	Country	Continent/ Region
CONTENT	City	X	If city IN state	If city IN country	If city ON continent
	State	/	X	If state IN country	If state ON continent
	Country	/	/	X	If country ON continent
	Continent/ Region	/	/	/	X

Figure 12: Decision table on how location could be matched for personalisation

The concept has so far been discussed amongst the user partners and has been refined. Still some open questions remain, with the main question being the most important one: Does this make sense for users and do they understand it?

This will have to be further evaluated in pilot 3 when the location-based recommendation has been discussed in detail with the technical partners and will be further implemented in the prototype.

The technical difficulty is still unclear (e.g. how to extrapolate the indicated locations from a list of imperfect entities?), and the right mix with other recommendation aspects has yet to be defined. A technical starting point could be to just focus on named places

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

and add the indirect mentions later. That will then also bring up questions of how to deal with "locations in between", like the segment of a highway (which would be interesting for traffic announcement) or regions (e.g. when it comes to weather forecasts) a user is on or passing through. A major task in this will be a clear mapping of what matches what.

There will be further difficulties in this process, as locations are sometimes used to refer to a person or organisation (like in "Berlin has so far refused these deals" - as in "The German Government"). Sometimes it's also the opposite: Cities that are not directly mentioned, but there is a clear relation to the news (e.g. Volkswagen announces letting people go, but Wolfsburg is not mentioned - still people there would be interested in the news).

The concept has not been realised in the CPN prototype and has hence not been evaluated any further. The consortium still needs to work on the finer details and discuss the proposed setup further both on a user and on a technical level.

3.7 FEEDBACK QUESTIONS

Perfect personalisation is a difficult thing. While it is easy to detect recurring patterns in media content a user consumes (such as certain entities like people or locations), the user action itself (e.g. reading, scrolling, dismissing, ignoring) is not always clear. Without questioning the motives of a user behind those actions, it's impossible to know what a user really means. This is why the consortium envisioned the option to set up automated questions for the user to be asked in irregular intervals, based on their actions to get further feedback.

The idea behind these feedback questions was to identify repeating patterns in the consumption of a user and learn more about the specific motivation of the users to behave in a certain way. In order to set up a logic like that, the system needs to be able to fulfill certain prerequisites, ranging from a simple feedback dialog pop-up to identifying entities in articles and patterns in the user behavior. It also needs clear defined thresholds regarding the trigger moments for the feedback questions and the actual, well-formulated questions and fitting answers in order to get useful results.

While all of this does not sound very complicated, the issue proved to be a lot more difficult both on a conceptual as well as a technical level. Conceptually, it was unclear how to best define aforementioned thresholds as well as how to best formulate questions and answers. On a technical level, the quality of the entities extracted from the news items as well as detecting the proper patterns in relation to them posed a much greater issue than was solvable in the short time available.

Since the idea was heavily debated in regards with whether it would be more helpful or more annoying to users, the consortium decided to put the feedback-question feature aside for the time being. Also, because focusing on the quality of the entities

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

first seemed more important, as they are also much needed for other features (e.g. transparency).

3.8 TRANSPARENCY

As an overall concept, transparency is a major feature for a successful personalisation application. "Why am I being shown this particular news item", is the core question a user asks while using such an app. And making this more obvious for the user was at the core of this deep dive.

As transparency as a feature had already been foreseen for the CPN app from the beginning, this wasn't planned as a feature deep-dive. But since the topic gained a lot more attention during year 2 through several incidents at some of the large social networks[13], it was decided to spend even more time on it. Additionally, the ongoing research in the topic and the development of the features for the CPN-app made it clear that there is a strong connection between good transparency (and how it is realized) and other core elements for a good personalization such as clear categories, keywords and extracted entities. Because the project is still working on the latter points, the transparency features currently available in the CPN application are still on an experimental level.

To better understand the options, the consortium used this feature deep dive to do an analysis of transparency features currently in place by major social media networks. DW took a closer look at Facebook, Twitter, Instagram and YouTube to see what they are doing to inform their users about how their suggestions of content and connections come about. The result in short is that there isn't a lot and the few things that are happening are not going very deep.

While Facebook, Instagram (the latter belongs to the first) and Twitter have at least some small features to give users some explanations (see Figure 13), YouTube doesn't have any explanations at all. Of course, it is sometimes obvious why users are shown a specific item, simply because we all know what we've clicked on, who we follow or what we've consumed before. But when it comes to the items further down the list and the particular order they appear in or the advertisements we are shown, not knowing how this came to be can be frightening. Applying these measures to news and not just entertainment, it can quickly become a real problem interfering with a well-informed public using these channels.

[13] https://www.nytimes.com/2018/09/28/technology/facebook-hack-data-breach.html

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

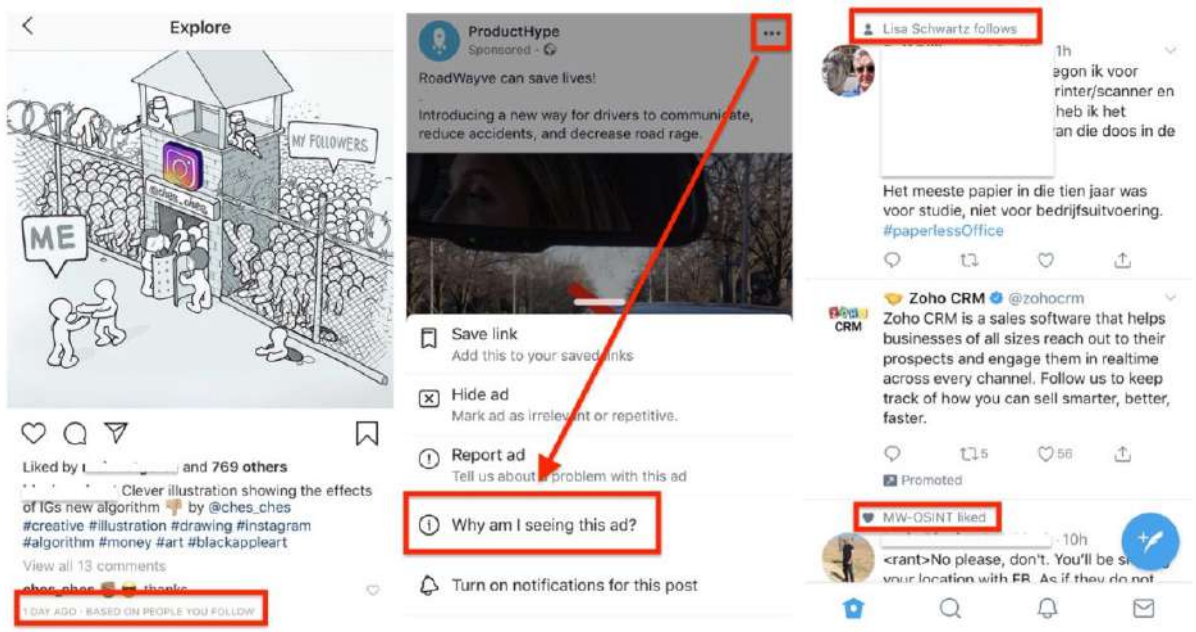


Figure 13: Examples of simple transparency features as they can be found on Instagram, Facebook or Twitter

Taking a closer look at the examples from Facebook, Instagram and Twitter, the first clear problem is that some of the transparency measures are simply not very transparent. Sometimes they are hidden and you have to know where to look. Sometimes they are simply too short, only giving a superficial explanation. And in many cases the details are not very easy to understand or again, too short, to really explain how the system works.

All of these networks only explain things down to a certain level, not giving full access to the algorithm. This is understandable to a certain degree as a company like Facebook doesn't want to give away its secrets about something so essential like the algorithm triggering the data recommendation[14]. But one has to keep in mind that Facebook (and the others as well) started out as networks among friends and have now grown to world-spanning services, connecting literally billions of people. All the networks have taken on so many areas of our lives that their influence on our lives has become enormous. With this comes a great responsibility, which most of the social networks have yet to really take on and deal with responsibly. Again, for the news sector this would mean to make people aware why they are seeing a specific item and not another one, as well as giving them the chance to actually have a look at the other side of things as well.

[14] https://wallaroomedia.com/facebook-newsfeed-algorithm-history/

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

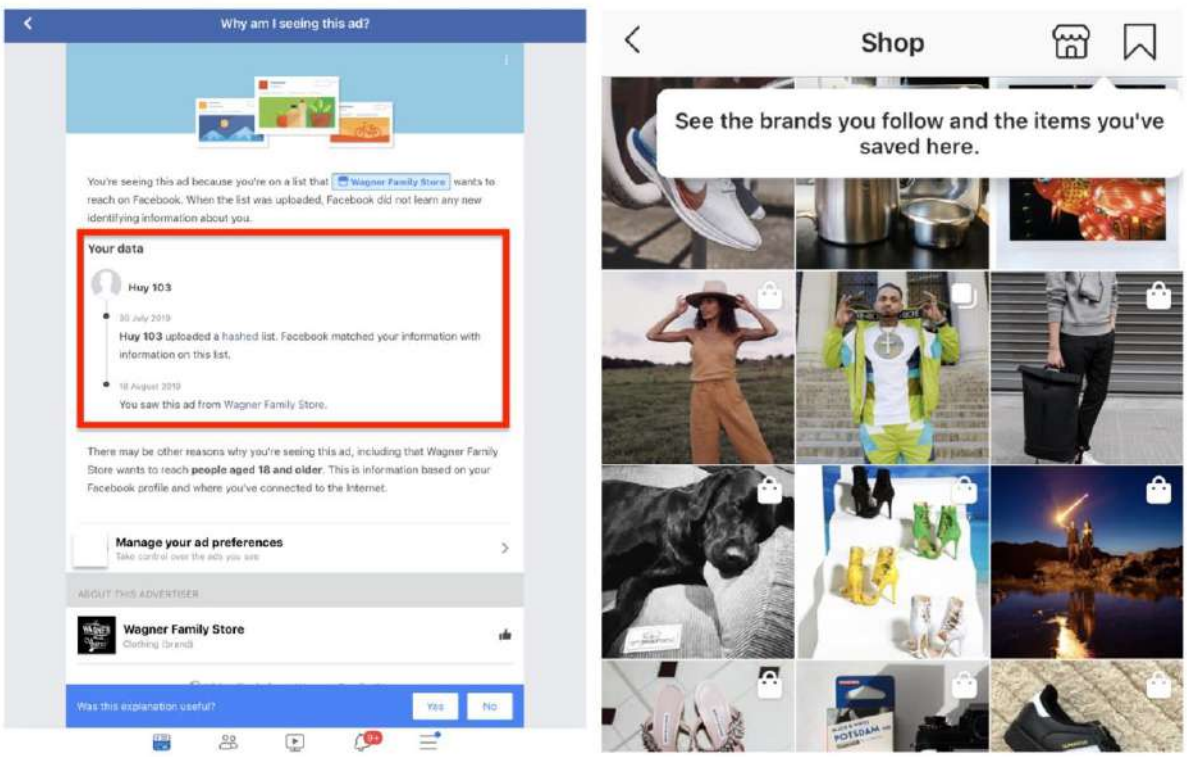


Figure 14: Examples of further levels of details on Facebook and Instagram

Through the research and further expert interviews the consortium did, it became clear that transparency has to work on two levels. There is a superficial one, that explains the "obvious" connection between a recommended item and the user's interest ("because you've read this, you get that") and an in-depth one, that goes into the details of the algorithm explaining how exactly a decision to recommend a specific item came to be. While the first one can be displayed in many different ways, the second one is a lot more complicated to bring across. For one, you have to find the right place and format to actually display all the necessary information, while for another, you need to find a way to make highly complex algorithmic structures and processes understandable to non-experts.

As part of its "transparency from the start" vision, the CPN consortium did include first transparency measures in the app from the beginning. This means for example, that users can access an overview of what news they had already consumed. The pilot evaluation later showed that this feature was not well placed and users weren't aware of it. What CPN needs is something similar to what Apple does with its screen-time feature (see chapter on FOMO) and general weekly reminders.

For the second pilot, an additional info button was added to each article, allowing users to get a short explanation on why this item was shown to them (see Figure 15). The texts are as of now still very generic and need to be both improved as well as connected to the extracted entities. The ideal case would be to clearly tell users that because of a person they seem to have a major interest in, this news item was chosen and places here and now in their timeline. The consortium used this feature deep dive

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

to further discuss this. The actual results are dependent on the improvements of the entity extraction and categorization efforts during year 3.

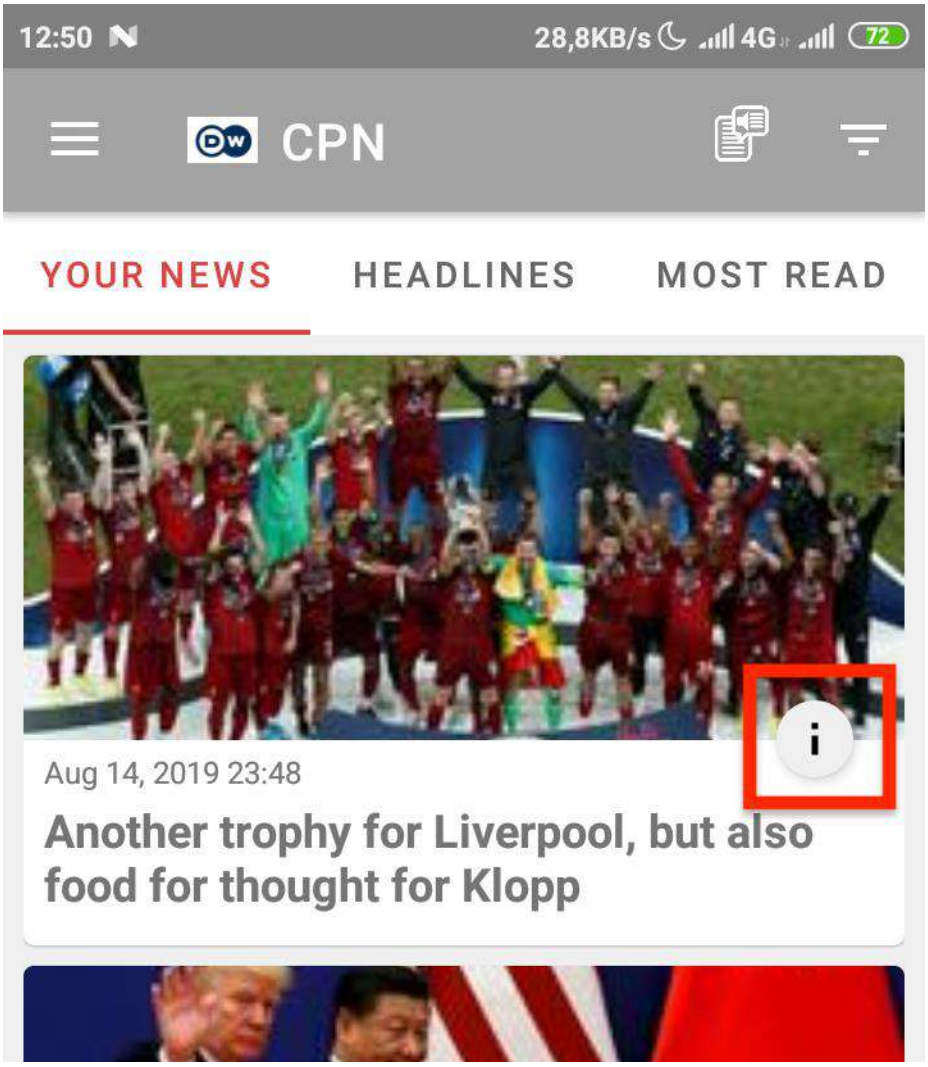


Figure 15: Example of info-button available on the CPN app prototype to learn more about the recommendation

Both of these measures however are only on the aforementioned shallow level. They give users an indication as to why something was shown and let them know what they already read. But they do not explain the workings of the algorithm in the background. Members of the CPN consortium sat down with Gordon Edall, the Director of Globe Labs at The Globe and Mail, Canada and had an in-depth discussion on the topic. The Globe Lab is deeply involved in projects around data science, natural language processing and predictive and advanced analytics, with Gordon being an expert in this field. The discussion showed that making the inner workings of the CPN algorithm visible and understandable isn't an easy task and the consortium decided to postpone further works on this for year 3. A first step, in order to enhance the transparency on this level is to take a closer look at the results from the Pilot 2 user tests and see what

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

people thought about the current features, before asking them about more in depth solutions and their actual requirements on this.

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

4 SIDETRACKS

The sidetracks were set up by the media partners in close collaboration with the technical partners. The initial idea was to explore some ideas and features, that couldn't be directly implemented in the application. During the process this goal changed to also focus on (pre-)testing the recommender system in an external environment, which was not part of the CPN platform, but was either already in production as the case for the DW sidetrack and the Dias sidetrack or created and optimized in-house by a media partner, as in the case of the VRT sidetrack.

LIVETECH and VRT both set up a recommender system for use in the VRT app, testing both the overall acceptance with its core audience as well as several specific features.

The technical set up of the sidetracks as well as the integration process, was already described in detail in deliverable D2.3. In this chapter, the focus is on the evaluation process and the results from a user partner's point of view.

4.1 NEWS APP (VRT)

The CPN platform was envisioned as a modular service and for a successful and client-oriented development, as we understand that our target group would be interested in enhancing existing applications, which are already familiar to their audience we needed to find out if the partial implementation of various CPN modules could be a feasible concept.

The VRT sidetrack had three main objectives:

- To find the limitations when implementing CPN components in a native news system like VRT's
- To test the recommendation engine and its effectiveness outside CPN
- To test acceptance of a personalized news offer with a known audience

Recruitment of the Participants for the application was done by means of a news article that was posted on the official VRT NWS website as well as the VRT Innovation facebook page. In this article, the concept of personalisation was explained. It also explained that VRT was looking for testers to test a personalised VRT NWS test app. About 1200 interested readers subscribed for testing after reading that article.

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

Figure 16: Screenshot of the VRT New App, using the CPN recommendation

4.1.1 User Behavior Tracking

In order to test this possibility, we built a news app that is in many aspects similar to VRT's main news app, but has some simplifications: it is optimized for mobile use, and the landing page for registered users has three streams of news articles:

- "MIJN NIEUWS" contains recommendation and is the stream where we tried different personalization strategies. The other two streams are the same for every user, and identical to the equivalent streams on VRT's main news website (latency effects ignored).
- "HOOFDPUNTEN" shows a selection of articles by the newsroom
- "NET BINNEN" gives the most recent articles in order of publication (latest first)

We tried the two personalization strategies in two phases of two weeks each. The users have been onboarded to the one-month long test and have received motivational emails to keep using the application regularly.

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

Phase 1

The list of recommended articles in the "MIJN NIEUWS" stream is a mix of lists generated by three algorithms:

- Collaborative filtering (CF)
- Content-based recommender (CB)
- Random selection of articles of the last n days (RND)

The random selection of articles is added to the recommendation list in order to avoid giving users too much of the same. CB and CF are often said to lead users into a filter bubble**.**

The final recommendation list is composed by adding different fractions of articles from the three algorithms. For example, for fractions (0.5, 0.3, 0.2), the list contains 50% percent of articles generated by CF, 30% by CB and 20% by RND. These fractions are tunable parameters of the overall recommendation algorithm. For example, increasing the last fraction would probably lead to a more diverse and surprising list. Setting it very high, however, would make the user feel like this list has nothing of interest to her.

The order of articles in the list is based on scores calculated by the recommender algorithms. These scores are measures for how well the recommended articles fit to the user's interests.

We divided the users in two randomly chosen groups, and attributed the following parameter sets:

Table 2: Recommendation mix ratio per VRT user group

	Collaborative filtering	Content-based recommender	Random selection of articles
Group A	33%	34%	33%
Group B	40%	34%	20%

Unfortunately, back-end problems caused the recommended articles to drift away in time. The later during the test period, the older the recommended articles were. That is very problematic, because barely nobody is interested in news articles older than two days. By the end of the two-week period, the lists in the "MIJN NIEUWS" stream contained only articles that were at least eight days old. This article age kept on growing every day.

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

Phase 2

One of the easiest and cheapest recommenders is the popularity recommender (POP). The recommendation list here consists purely of the most popular articles of the last hours or days. Notice that this list is the same for every user, and that we cannot strictly speak of personalization.

POP is very robust: if you want to recommend the same list to every user, POP contains the articles that have the highest probability of being liked. In the recommendation community, POP is often used as a baseline recommender: if your system cannot achieve better results than POP, it is not ready for large-scale deployment.

POP, as described here, has one tunable parameter: the period of time over which the most popular articles are calculated. We wondered what would be the best time period: is it a couple of hours, or several days?

We divided the users in three randomly chosen groups, and attributed the following parameter values:

Table 3: Different time length for the calculation of most popular items per test group

Group	Time spent
1	1 hour
2	12 hours
3	2 days

The list of popular articles was calculated based on numbers from VRT's main news website and app. This platform has several hundreds of thousands of users daily, and hence give very stable results. Calculating popular articles on a separate website instead of on the test also removes unwanted influences between the different groups.

Group 1 gets articles that were trending during the most recent period, group 3 during a much longer period. As a result, the "MIJN NIEUWS" stream for group 1 contains usually more recently published articles. However, an article in this list can be one that is trending only for a very short period and may in the long run turn out not to be that important for the audience. On the other hand, group 3 also gets also articles that were trending during the day before. These articles are a bit older and possible not interesting anymore. Regular users also might have already read many of these articles.

Group 2 could be a compromise between the two. VRT's main news app has a stream "MEEST GELEZEN" (MOST READ), that contains exactly what is in the "MIJN NIEUWS" stream of group 2 (latency effects ignored). Interestingly, the corresponding tab in

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

VRT's news app uses a 2-hour period. Hence, this test can also serve as an evaluation of the main app's strategy for this stream.

We did no filtering on the age of the articles. The popularity recommender is robust enough to select mainly recent articles. At a certain moment, we saw an article that was six months old, but still was recommended by the popularity recommender. This article was a slightly modified, older article on the debate on daylight saving time, a topic that was relevant at that moment, as western Europe was changing clocks.

4.1.2 User Feedback

We sent out two short surveys to get a clearer view on how the end user perceived the MyNWS experience. The first survey, in which 146 respondents took part, was sent out after 2 weeks, i.e. right after phase 1, see introduction. Following phase 2 we sent out a closing survey, on which 292 users responded. We asked a few questions from the previous survey again, so that we were able to compare. The participants were fairly distributed amongst age groups. Around 70% of the participants in Phase 1 were male.

"How informed do you feel after reading the news on the tab 'My News' on a scale from 1 to 10?" 1= 'I don't feel at all informed.' 10= 'I Feel very informed.'

To what extent do you agree with these statements (Totally disagree to completely agree)

I have the feeling that the news has been offered especially for me?

I have the feeling I have been missing out on something.

I wish to receive articles outside my field of interest in the tab 'My News' to keep me completely informed (only Survey 1)

There was an open-ended question about the age of the news articles (only Survey 1). At the end of the Survey 1 there were also some demographic questions, i.e., age, sex, occupation, education, location.

Question 1.

"How informed do you feel after reading the tab "My News"

We can conclude that following phase 2 participants had a higher score on our feeling of being informed scale compared to phase 1.

Question 2.

I have the feeling that the news has been offered especially for me?

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

We see in phase 1 that less users agreed with the feeling the news was personalised compared to Phase 2, which might indicate that the recommender did not succeed in offering a personalized news experience to users.

I have the feeling I have been missing out on something.

In phase 2, less people felt they were missing out on something compared to phase 1. This could mean that the first recommender was filtering too many articles, giving readers the feeling that they were not fully informed.

I wish to receive articles outside my field of interest in the tab My News to keep me completely informed (only Survey 1) In addition to their personalized articles, 75% of the first survey's respondents indicated they also want articles outside their field of interest in MyNWS.

Question 3.

Age of the news articles.

The more the test period in phase 1 progressed, the older the recommended articles seemed to be. This is not what people want. This was the most mentioned remark in the feedback, 44 times. Via the first survey we asked users how old a news item should be in the list of recommended articles.

Opinions differed from half a day to 1 week, but most of the respondents felt that articles should be maximum 2 days old. Some respondents mentioned that articles could be older, especially when the articles are personalised, but they would prefer them to be in chronological order.

There were users that preferred the articles to last longer than two days. These users didn't consume the news every day and didn't want to miss out. Some suggested to add an option so users could choose for themselves.

There were also some people that couldn't give a clear answer because they think the maximal allowable age of an article should depend on what type of item it is. For an article about an important event that has been the main point of the news, e.g. Brexit, it is not appropriate to show it again the next day. For articles about culture, media, etc., an article can last longer. E.g: the article about the famous architect Le Corbusier today could still be relevant a week later. People suggest that this last type of newsitem could be in another tab, e.g. archive. We finished the closing survey by asking the respondents to give us tips on how to further develop MyNWS. We received a lot of comments about the age of the news items. Another much mentioned remark was that people didn't feel completely informed, they used other news channels to be sure they didn't t miss out on anything.

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

4.1.3 Results

Results operational data

Recommendation systems that generate lists of too old articles, have a detrimental effect on the usage. Article recency is very important.

The activity of the users decreased throughout the four-week period. Many users could be reactivated by motivation emails. However, the ratio of clicks in the personalized stream was much higher during phase 2 (popularity recommenders) than during phase 1 (hybrid recommenders). Calculating the recommendations of the popularity recommender on the last one hour of data leads to more clicks than on the last twelve hours. Calculating POP on the last two days results in lower average click ratio in the personalized news stream than calculating POP on the last hour or the last 12 hours. More detailed operational results can be found in the appendix.

Combining Operational data with Experience data (data from the app with survey data)

This experiment covers usage data from the app, called operational data (O-data), i.e., this is the click data, which article they read, when, and from which stream, etc.

On the other hand, this experiment includes data from surveys as well, called experience data (X-data). This is their subjective feeling of being informed, having read personalized news, feeling of missing out.

If the ultimate goal of our news media is to serve the citizens with the best news information, i.e. to inform the citizens as good as possible in current affairs, a personalized news offer would be able to increase their feeling of being informed. In the current information tsunamis, it is increasingly harder to make the best of your time to find the right info you are looking for. See the "Time-Well-Spend" movement increasing their voice on this issue. Hence, these hypotheses about the combined data could be made: Expectation intuitively: The more people read in general, the more informed they feel.

The more people read the personalized MyNWS stream, the more they feel informed after reading the MyNWS stream.

Hence, ideally, the higher the click rate on the personalised stream, the higher they rate their feeling of being informed. Our analysis concludes that unfortunately there are no meaningful correlations between click ratio in the MyNWS stream and the feeling of being informed in any group. Overall, we found that the recency of the articles is of utmost importance in news recommender systems. Recommending too many old articles is the largest problem in this test. A popularity recommender is a first step towards a stable and reliable system. However, this recommender offers no real personalisation. Popularity recommendations can best be calculated on a short period of time (here: one hour). There were no correlations between the survey results and the user behavior.

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

4.2 SMART TV INTEGRATION (DW)

DW has planned a sidetrack, in the form of a test integration of the CPN recommendation engine into its existing Android Smart TV application and Fire Smart TV application. The Smart TV applications are a DW niche product with a few thousand users a month. The integration will use an additional content row in both Smart TV applications, which will consist of content items recommended by the CPN algorithm.

Figure 17: Screenshot of the current DW SmartTV app setup (without recommendation)

Similar to the VRT track, the goal of this implementation is to test a potential use case of a media company partially integrating components or services of the CPN platform into existing applications. This way, it's possible to add innovative functionality without overwhelming users with a completely new application or interface. Furthermore, we also want to compare the performance of the CPN recommendation engine against a keyword-based content personalization solution, that was created at DW in-house, without the involvement of CPN.

The goal is to see which recommendation engine is performing better in retaining audience. For this purpose, we will measure and compare specific KPIs, such as time spent in the app, the number of consumed content items and time spent with each content item. The DW sidetrack has been scheduled after Pilot 2 and the VRT-sidetrack, but before pilot 3 for early fall 2019 with the overall duration of one month. That way it would not interfere with other developments. Due to technical challenges the realisation could not be finalized as planned in year 2.

The test group will include about 1000 existent Smart TV Beta users. The test-users will not get any specific instructions, but will continue using the application in the way

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

they usually do. This approach is consistent with the usual Smart TV Beta testing. Within the duration of a month an existing content row will showcase content selected by the CPN recommendation engine to one group of users and content selected through a recommendation approach created in-house to another group of users.

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

5 PILOT 2 EVALUATION

As described in the proposal, one of the objectives of CPN is "to iteratively test and validate the solution in (near-to) operational real-life environments in different countries, Belgium, Germany and Cyprus; by implementing large-scale pilots" (CPN proposal, p. 5).

A co-creation process is applied to iteratively test and develop the CPN solution. This will be done in three different stages, including a larger number of users as the maturity of the developed proof of concept increases. The first phase (pilot 1) was the controlled lab phase, involving approximately 50 end- users in Belgium and 20 friendly users in Germany and Cyprus end 2018. The second phase (Pilot 2) was the closed group testing and was organized in May-June 2019. The set-up and results of Pilot 2 are described in this chapter. In the last research cycle, an open living lab phase will take place, meaning that everyone can participate in this pilot phase. The aim for this live beta phase is to have 300 participants involved in Belgium and Germany and 200 in Cyprus.

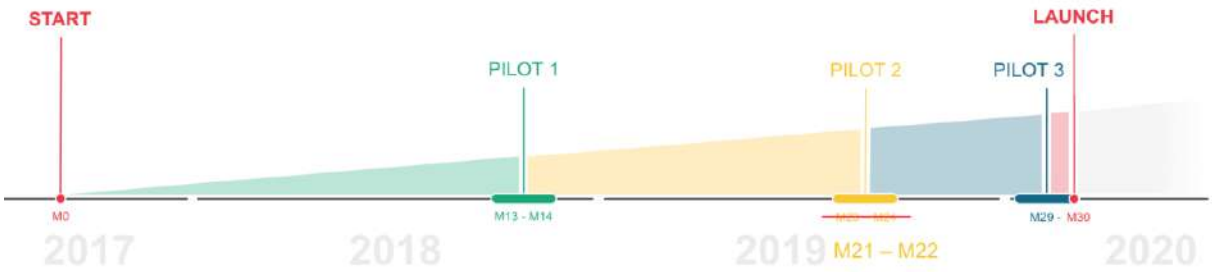


Figure 18: Overview of project timeplan in regards to pilot evaluations

The following table shows the original planned user involvement in each pilot phase. For Pilot 2 there were fewer end-users involved than originally planned. This has several reasons. A first reason is that in Belgium, xx users were already involved in the VRT sidetrack, leading to relevant insights for the Pilot 2 related- questions as well, which made the need for additional respondents less stringent. Second, the timing of the test (June) isn't the easiest period to recruit participants and as we asked for an intensive participation during 4 weeks, this is difficult to commit to for some people. Third, also the requirement for a recent Android device to use the app made that several potential participants couldn't participate. Fourth, the CPN recommender wasn't functioning as well as expected in this phase of the project, leading to a need for fewer participants to get relevant feedback than typically planned in this pilot phase.

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

Because of this, we will increase our participant numbers in the third and open pilot phase and invest in new ways to recruit potential participants (e.g. via the media partners channels, social media etc.)

Table 4: Overview of planned participant numbers per media partner for all three CPN pilots

	VRT	DW	Dias
Pilot 1	50 end-users	20 friendly users	20 friendly users
Pilot 2	200 end-users	150 end-users	150 end-users
Pilot 3	Maximum 300 Open testing	Maximum 300 Open Testing	Maximum 250 Open Testing

5.1 METHODOLOGY

During Pilot 2, the app was tested by end-users of the three user-partner media organisations simultaneously. The goal was to get end-user feedback on the developed recommender and to evaluate the algorithm behind it. The main research questions for Pilot 2 were:

- Does the recommender work? (From a technical and usability point of view)
- How do end-users make use of the news recommender?
- How do end-users evaluate the personalised news they receive?

Additionally, in the interviews, focus was placed upon Fear of Missing out (FOMO), feedback mechanisms, and the uplifting/depressing article classifier.

The Pilot 2 end-user evaluation consisted of 4 research actions:

- ➔ Research action 1: The zero-measurement survey
- ➔ Research action 2: Testing the recommender system
- ➔ Research action 3: The follow-up survey
- ➔ Research action 4: Interview or focus group discussion

These research actions were executed chronologically in each pilot country. The figure below shows the research process. These actions are described in more detail in the following sections.

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

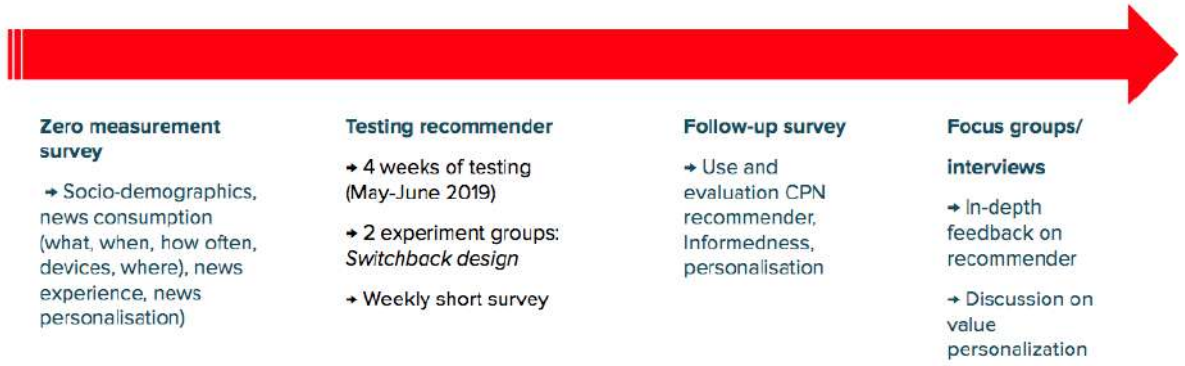


Figure 19: Overview of the order of the different measurement methods applied during pilot 3 evaluation

Research action 1: Zero measurement survey

First, end-users could subscribe to the pilot by filling out a zero-measurement survey. This survey allowed the consortium to build participant profiles to know what type of news consumers the participants are.

An online survey was created using the Qualtrics survey software. In Belgium, for the VRT pilot, the survey was distributed in Dutch. In Germany, for the Deutsche Welle (DW) pilot, the survey was distributed in English and in Cyprus, for the DIAS pilot, the survey was in Greek. Appendix C shows the survey questions.

88 respondents completed the survey for the VRT pilot, 78 respondents completed the survey in DW pilot and 29 for the DIAS pilot.

Research action 2: Testing the recommender

Participants were asked to use the application at least once a day for 4 weeks, from Monday May 20, 2019 to Sunday June 16, 2019. They were asked to download the application on their own Android smartphone. The application was available in the Google play store.

We used a **switchback design** in the Pilot 2 test. The users were randomly divided in **test group A** and **control group B.** Group A received personalised content in the first week, no personalisation in the second week, personalised content in the third week, and so on. Group B followed the opposite trajectory. When personalisation is off, the PERSONALISED stream is filled with the most popular items at that time. The advantages of this switchback design are: we can compare usage and survey feedback between users that receive personalised content and those who don't. Every user gets a taste of our personalisation system (half of the time).

Imbalances between groups are cancelled out. Suppose that, by accident, group A has users that on average consume more news than group B. A non-switching design,

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

where group A always receives personalised content and group B never, would overestimate the consumption due to personalisation. Additionally, we used a number of qualitative surveys before, during and after the test. We started with a pre-measurement survey and continued to send out short surveys after very switch back during the testing. After finalising the 4 weeks of live testing, the participants have been given an extended survey. The final evaluation part was a focus group conducted with a few selected participants.

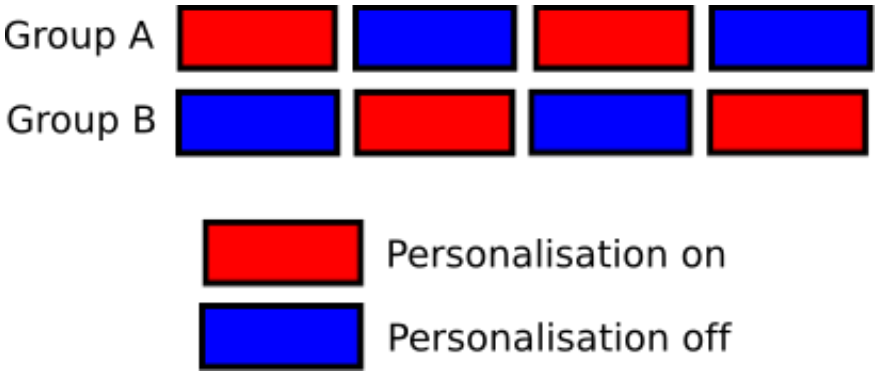


Figure 20: Visual explanation of the switchback theme applied to the two groups evaluating the CPN Pilot 2

Every Monday night, the two groups were switched.

The participants were informed on the switchback design before the start of the test in their test instructions, but they didn't know in which group they were divided.

Every Monday, a short online survey was sent to the respondents in order to assess their experience with the personalized news stream in the past week and compare the responses between the control and test group, and within groups over the weeks. The following 2 questions were asked every week in these **weekly surveys**:

- ***"How informed do you feel after reading the news on the tab 'My News' on a scale from 0 to 10?" 0= 'I don't feel at all informed.' 10= 'I feel very informed.'***
- ***"To what extent do you agree with these statements." (5-point scale: Totally disagree to completely agree)***
- ***I have the feeling that the news has been offered especially for me.***
- ***I have the feeling I have been missing out on something.***

The table below shows the number of participants in each group.

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

Table 5: Overview of actual participant numbers per media partner in the Pilot 2 evaluation

	Group A	Group B
DIAS	17	14
DW	25	24
VRT	35	35

The table below gives an overview of the Pilot 2 timeline:

Table 6: Timeline of the Pilot 2 Evaluation

Monday 20/05	Start of testing period
Monday 27/05 14:00	Weekly survey
Tuesday 28/05 03:00	Switch
Monday 03/06 14:00	Weekly survey
Tuesday 04/06 03:00	Switch
Monday 10/06 14:00	Weekly survey
Tuesday 11/06 03:00	Switch
Monday 17/06 14:00	End of testing period Final survey

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

The advantages of this switchback design are:

- ➔ We can compare usage and survey feedback between users that receive personalised content and those who don't.
- ➔ Every user gets a taste of our personalisation system (half of the time).
- ➔ Imbalances between groups are canceled out. Suppose that, by accident, group A has users that on average consume more news than group B. A non-switching design, where group A always receives personalised content and group B never, would overestimate the consumption due to personalisation.

The disadvantages of this switchback design are:

- ➔ Users might feel the app is not working properly, depending on which test-group they are in.
- ➔ Some participant did not read the instructions well and weren't aware of the fact that it is was an experimental design.

Research action 3: Post-test Survey

After finalizing the 4 weeks of live test the participants have been given an extended survey.

At the end of Pilot 2, after 4 weeks of testing, participants were asked to complete a final survey. The goal of this second survey was to evaluate the CPN application. (The full survey can be found in the appendix)

The survey was completed by 57 participants of the VRT pilot, 15 of the DW pilot and 13 of the DIAS pilot.

Research action 4: Interviews and focus group discussions

The final evaluation part consisted of interviews with a few selected participants.

At the end of Pilot 2, in June/July 2019, we invited the participants of the VRT and DIAS pilot to take part in an interview. Through these discussions, we wanted to get more in-depth feedback on the recommender. Furthermore, additional questions were asked on the value of personalization.

In Belgium, 7 participants were interviewed (1 group interview with 3 people + 4 individual interviews) and in Cyprus, 9 people were interviewed (9 individual interviews). All interviews were conducted online (with a conference call tool) or over the telephone call.

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

All interviews were conducted online (with a conference call tool) or over the telephone. The interviews had an average duration of 30 tot 60 minutes and discussed the following topics:

- Evaluation of the application: first impressions, positive, negative, missing features
- News personalization: Interest in news personalization, the personalized news feed in the CPN app, Fear Of Missing Out
- Feedback mechanisms
- First opinions on uplifting/depressing article classifier

DW did not set up a focus group because the participants all live in different time zones and therefore a focus group, even as an online teleconference version would not be feasible. Instead, DW will look into the possibility to set up small local focus groups at a later time. The problem here is that a local focus group would not necessarily represent DW's international target audience.

In the appendix X, the full topic list can be found. All participants signed an informed consent form prior to their participation in the focus interviews.

5.2 PARTICIPANTS RECRUITMENT REPORT

The following subchapter will give an overview of the user recruitment for Pilot 2 in via VRT (Dutch), DIAS (Greek) and Deutsche Welle (English). The recruitment was undertaken independently by all three media companies.

5.2.1 VRT

In Belgium, for the VRT pilot, users were recruited for Pilot 2 via several means. First, the participants of pilot 1 who indicated they could be contacted again for Pilot 2, were invited to participate. Further, people were recruited via personal and professional networks and social media (Facebook, Twitter, Linkedin). A visual was created that could be easily shared via social media and colleagues of VRT and IMEC employees were asked to spread the message.

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

- A sponsored post was launched on Facebook as well to increase participation. The add was online for 10 days and recruitment criteria were:
- Lives in Flanders
- Speaks Dutch
- Age: 35+
- Has Android phone

The post reached almost 1000 people, but unfortunately didn't lead to many extra subscribers. This is probably due to the fact that 'CPN' is currently not a known or trusted news source.

Figure 21: IMEC Recruitment Post to reach out for possible participants

5.2.2 DW

As already done in year 1, DW reached out to potential users through its customer service department to reach out to possible users. For the user surveys done in relation to the requirements gathering at the beginning of the project, this had worked quite

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

well. At the end of the survey, participants had been asked if they were interested in participating in further research. All positive responses were noted in an extra list.

As the response rate of those on this list was very low, a second round of recruitment was undertaken. In this second round, individuals who had not participated in the survey were included, as well. In total, almost 28,000 people were invited to participate in Pilot 2. All those individuals received an e-mail with an invitation as well as information about the app and the CPN project.

However, the response rate for Pilot 2 did not correspond with the targeted number of participants, only few people were willing to participate in the evaluation period. DW had registered 50 active users for Pilot 2 of which 15 completed all evaluation activities, including the final survey.

The results for this are not all clear. But since DW audience is a global audience, scattered across the globe, in different circumstances, it might have been the technical requirements as well as the length of the activity. It is difficult to keep people motivated only via mail over such a long time with no personal contact and no clear benefit for them.

For Pilot 3 DW will undertake several steps to overcome this issue. Alternative recruitment methods could be an option, as well as different, more qualitative and possibly shorter testing approaches.

Overall, this could result in an individual test design for DW which would differ from the other content partners.

5.2.3 DIAS

As recruitment measure DIAS published an article in the online magazine Sigmalive, titled "I want upgraded information - CPN"[15]. The article described the common problem that the users face due to lack of time during news consumption, ending up with the solution to this problem coming from CPN. There was also a presentation of all the partners of the project. The article included the links to the official site of CPN and to the video "Make your news stories stand out with CPN" on Youtube.

Also, the link to the Mailchimp[16] for subscription appeared three times in the text. The article, during its lifetime had 1,513 unique page views and a total of 1,619 views. (For details see the "Dias Performance Report 1" in the annex).

Afterwards, the article was posted on Facebook on 4/4/2019. It was a sponsored post from 5/4/2019 until 11/4/2019. It was promoted again on 22 - 26/4/2019 (total

15 http://www.sigmalive.com/news/local/561900/thelo-ekseligmeni-enimerosi-cpn

16 https://mailchimp.com

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

duration of promotion 10 days). The article[17] reached 40,757 users in total and had 1235 clicks. (For details see the "Dias Performance Report 2" in the annex).

Figure 22: Screenshot of the post DIAS used to recruit participants for the CPN Pilot 2 evaluation

On top of that Dias also presented the CPN project at the Google News Lab which was organized by Sigmalive and Google News Initiative on 07/05/2019, a workshop for journalists and media specialists. Christos Danezis, on his speech at the beginning of

17 https://bit.ly/30tWmSw

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

the workshop, made a brief presentation of CPN project, inviting the audience to participate in Pilot 2.

Outside the workshop, there was an information desk for those who wanted more info about the project. They also had the chance to subscribe to the CPN on the spot by completing the form online. The event was hosted at Sigma TV's main news broadcast on 08/05/2019. (Images can be found in the annex).

At the end of the recruitment process DIAS managed to have 146 subscribers on Mailchimp, 12 of them unsubscribed during the testing period.

5.2.4 Profile of the participants

The zero-measurement survey gave us some background information on the socio-demographic profile and news profile of the participants.

In total, about two-thirds of the sample (68,2%) is male. The table below shows the exact number for each pilot. For the VRT pilot, the male-female distribution is quite equal (53,4% male vs. 46,6% female participants), while for the DW and DIAS pilot there are around 80% male participants.

Table 7: Overall profile of the participants per media partner

		VRT	DW	DIAS	Total
Male	N	47	63	23	133
	%	53,40%.	80,80%.	79,30%	68,20%
Female	N	41	15	6	62
	%	46,60%	19,20%.	20,70%	31,80 %
Total	N	88	78	29	195
	%	100,00%.	100,00%	100,00%	100,00%

The participants in the VRT pilot are all living in Belgium and the participants of the DIAS pilot are all living in Cyprus. The participants of the DW pilot however, are living in 60 different countries.

79,5% claim to be very to extremely interested in news.

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

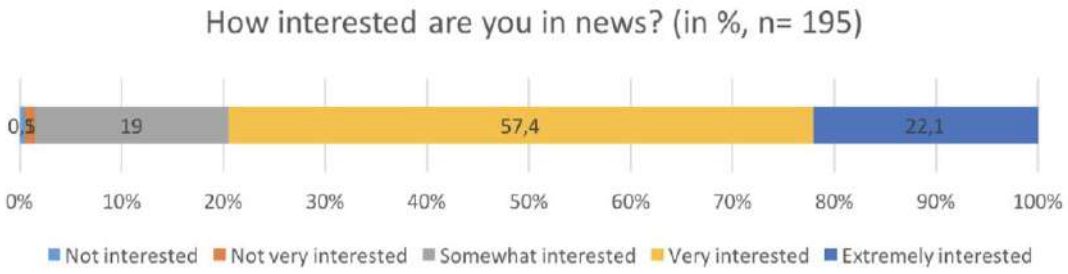


Figure 23: Results from the user survey: How interested are you in news?

The respondents were asked what their main reason was to follow the news. They were asked to rank the following 4 reasons in terms of importance:

- To stay informed about what is happening in the world
- To stay informed about what is happening in my country
- To stay informed about what is happening nearby
- To stay informed about specific topics such as sport, politics, showbiz, etc.

'To stay informed about what's happening in the world' was ranked of the highest importance as the number one reason by 56,1% of all respondents.

81,5% sometimes uses a news application, and **60,6% uses a news application at least once a day**.

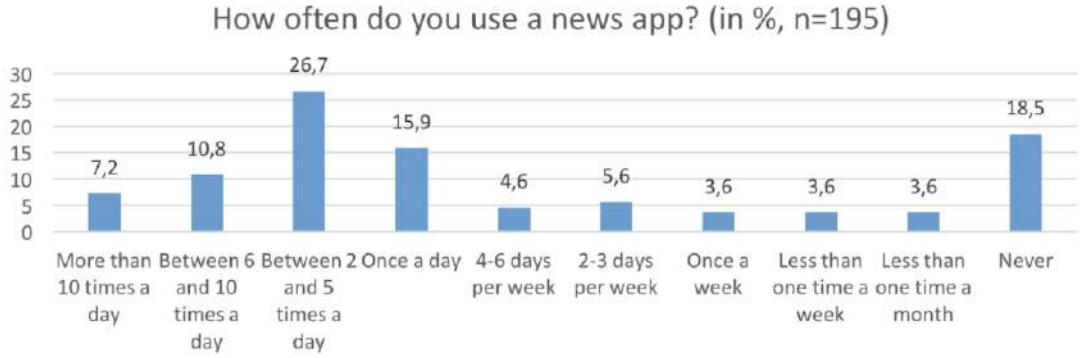


Figure 24: Results from the user survey: How often do you use a news app?

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

5.3 EVALUATION RESULTS

This chapter will provide an overview of the evaluation results. As mentioned above we measured quantitative KPI as well as qualitative data in form of user surveys. We will first list the qualitative KPI and then a summary of the survey results.

In a final step we will link the qualitative and quantitative data and discuss the results for all three media organisations.

5.3.1 Quantitative KPIs

In this section we will present the quantitative results of the Pilot.

Click history

The users in this pilot were highly motivated and opened many articles. Most of them stayed interested until the end of the four-week period. The figure below shows the daily clicks per channel. Blue bar represents weekdays, red bars are the weekends.

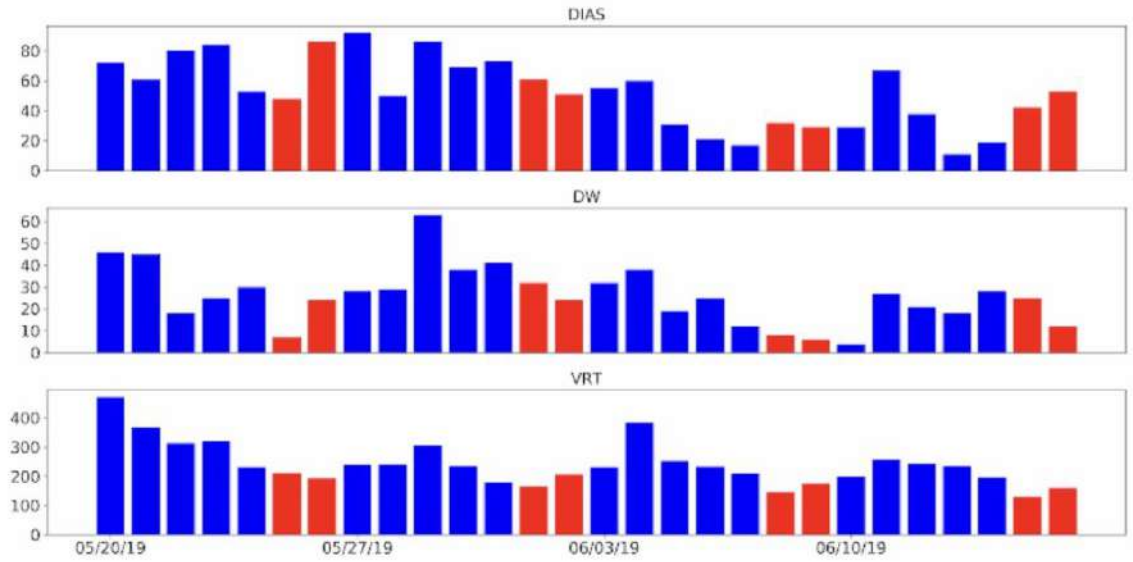


Figure 25: Overview of how often the users clicked on articles during the evaluation

Comparison between groups

Unless otherwise indicated, we use the Wilcoxon rank-sums p-value to evaluate the statistical significance of the difference between two lists of metrics, calculated per

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

user. For these tests, we yield two lists of the metric of interest, each with one entry per user.

Table 8: Users per group and per channel

Content Partner	Group A	Group B
Dias	17	14
DW	24	25
VRT	35	35

Clicks

We aggregated the events per user and divided these over periods where the user was receiving personalisation and when not. We counted the number of clicks in the subsets. This yields two lists, each with a number of clicks per user.

The table below shows the average number of clicks between events in the control and the test group, per channel.

Table 9: Average number of clicks between events in the control and the test group, per channel

	Dias	DW	VRT
Control Group	36	13	53
Test Group	24	13	57
p-Value	0.91	0.526	0.903

➔ The high p-values indicate that there are no significant differences.

Click fraction per stream

We calculated fraction of clicks in each stream, per user in each group, for all channels.

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

Table 10: Fraction of clicks in each stream, per user in each group, for all channels.

	Headlines	Personalised	Popular
Control Group	0.328	0.490	0.182
Test Group	0.306	0.525	0.186
P-value	0.422	0.710	0.598

➔ The high p-values indicate that there are no significant differences.

Reading time

We tracked the time a user spent on an article. This could be a great metric of interest: the longer a user reads an article, the more interested she probably finds it. The table below shows reading times in seconds.

Table 11: Reading time (in seconds) per media partner

	Dias	DW	VRT
Control Group	168	611	644
Test Group	167	604	388
p-value	0.206	0.857	0.189

➔ The high p-values indicate that there are no significant differences.

We noticed large outliers in the reading times. This might correspond to situations where the user was not actually reading anymore, but still had an article open.

We removed the reading times that were above the 95th percentile, and re-analysed. This 95th percentile lies at a reading time of 425 s.

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

Table 12: Reading time (in seconds) per media partner (cleaned)

	Dias	DW	VRT
Control Group	42	83	53
Test Group	39	88	58
p-Value	0.076	0.585	0.001

➔ The average values seem more realistic. For VRT, the low p-value indicates a small but significant difference between test and control group.

Scrolled full

We kept track of whether the users scrolled to the end of an article. Along with the reading time, this is a measure of the interest of the user in the content.

The value of this parameter is either 1 (full scroll) or 0 (not scrolled to the end of the article). It is hence a binomial distribution for each subset of events. Instead of the Wilcoxon rank-sums test, we use the McNemar p-value to check the statistical.

Table 13: Scroll-ratio of how often users scrolled to the end of an article

	Dias	DW	VRT
Control Group	0.683	0.651	0.594
Personalised Group	0.662	0.627	0.633
p-value	0.001	0.0	0.0

➔ The small p-values indicate significant differences between users that received personalization and those that did not. For DIAS and DW, the users that received personalization scrolled less often to the end of the articles, for VRT they did this more often.

Additional Metrics: As mentioned before we have included two additional metrics in the Evaluation.

Diversity

Whether the filter bubble actually exist is an ongoing discussion in the recommender community. We believe that the best way to learn more about this important topic is to test algorithms in realistic situations, and that is what we did during this pilot.

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

Suppose a user has read a number of articles during the test period, and another number of articles during the control period. A dangerous algorithm would yield a less diverse collection of articles during the test period than during the control period.

There is currently no generally accepted metric for the diversity of a collection of texts, hence we use one of the more straightforward metrics.

We constructed the tf-idf weighted bag-of-words matrix[18] for the articles read on VRT's channel. Every article is represented as a vector in this matrix, where the components are weighted counts of the words in the articles. We calculated the cosine distance between all pairs of articles. Articles that have few words in common, have a larger pairwise distance between them. For a collection of articles, we defined the diversity as the average distance between each pair of articles in the collection[19/20] This way, we obtained two lists of diversity metrics: a list for events when the user was in test period, and another one for when she was in control period.

The table below shows what these two lists look like (fictional numbers).

Table 14: Demo-table of diversity metrics between effect/no-effect

	Test period with effect	Test period without effect
User 1	0.432	0.568
User 2	0.563	0.339
…	…	…

➔ The Wilcoxon rank sums p-value was very high at 0.97, and we conclude that there was no difference in diversity. Hence, our recommender system does not cause users to consume less diverse news.

Emotional Make-up and effect of the news articles: Uplifting/depressing

The consortium has recently created an uplifting/depressing classifier for news articles. The classifier outputs three scores for an article: uplifting, neutral and depressing. These scores indicate how likely a user thinks the article belongs to each of these categories. In this pilot, this classifier has not been used to make recommendations.

18 D. Jurafsky and J. H. Martin, Speech and Language Processing, Third Edition draft, 2018, Stanford University

19 C.-H. Chiu and A. Chao, Distance-Based Functional Diversity Measures and Their Decomposition: A Framework Based on Hill Numbers, 2014, PLOS ONE, Volume 9, Issue 7

20 M. Kunaver and T. Pozˇrl, Diversity in Recommender Systems, A Survey, Knowledge-Based Systems (2017), doi: 10.1016/j.knosys.2017.02.009

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

However, it could be interesting to see if users consume more uplifting or more depressing articles in certain conditions.

We checked the difference between users in test or control group. The results can be found in the table below.

Table 15: Uplifting vs. depressing articles read by test users

	Depressing	Neutral	Uplifting
Control Group	0.261	0.528	0.211
Test Group	0.268	0.519	0.213
P-Value	0.090	0.025	0.440

➔ Only the Neutral score has a p-value below 0.05, however the difference between the means is small. We conclude that there is no difference in consumption of depressing, neutral or uplifting articles between users in test and control group.

5.3.2 Survey Evaluation

We conducted a number of surveys before and during the testing. We started with the zero-measurement surveys

Weekly surveys

The weekly surveys have been sent out to the participants by e-mail and covered these questions:

How informed do you feel after reading the news on the tab 'My News' on a scale from 0 to 10?" ***0= 'I don't feel at all informed.' 10= 'I feel very informed.'***

To what extent do you agree with these statements. (Totally disagree to completely agree)

I have the feeling that the news has been offered especially for me?

I have the feeling I have been missing out on something.

The below results are from VRT data only because the number of participants are too low for the other two news media companies due to not filling in the surveys to make meaningful comparisons.

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

Question 1: *"How informed do you feel after reading the news on the tab 'My News' on a scale from 0 to 10?"*

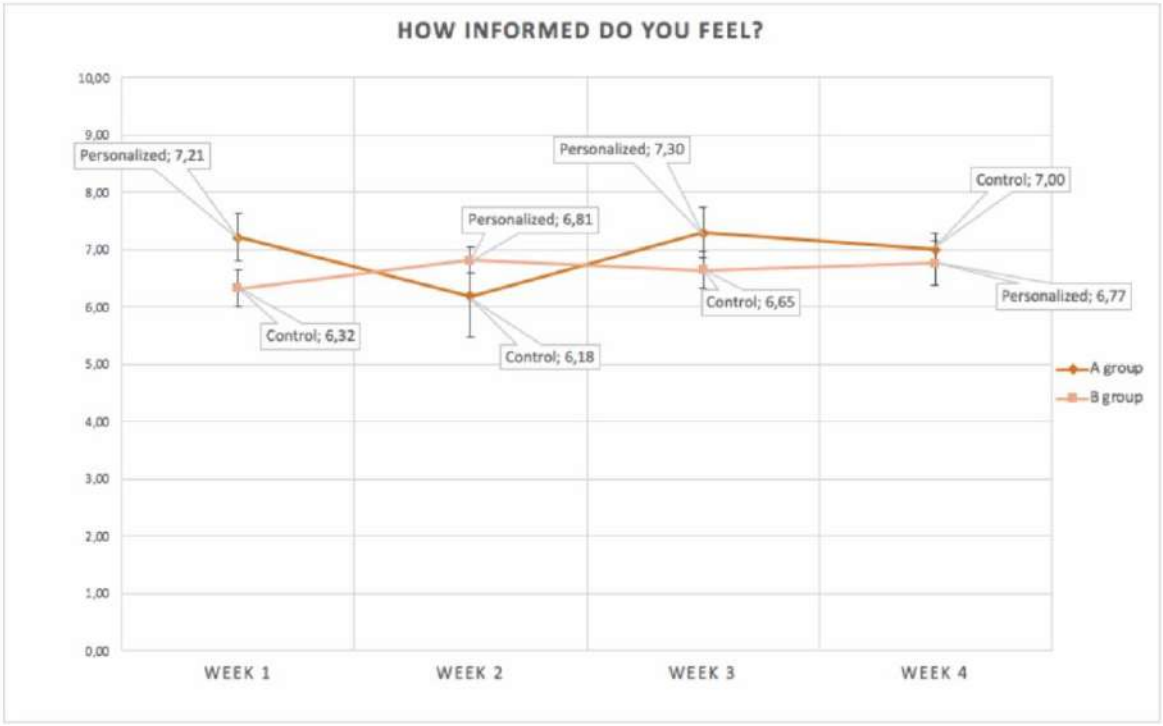


Figure 26: Means of group A and B over the four weeks. Error bars denote the standard error of the mean

➔ If we look for each week whether there is a difference between A and B group, we find one slight meaningful difference, namely in week 1 where the personalized group felt more informed. If we look at the results within a group and compare their scores per week we find no statistical difference.

(Detailed analyses can be found in the Annex.)

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

Question 2: *I have the feeling that the news has been offered especially for me?*

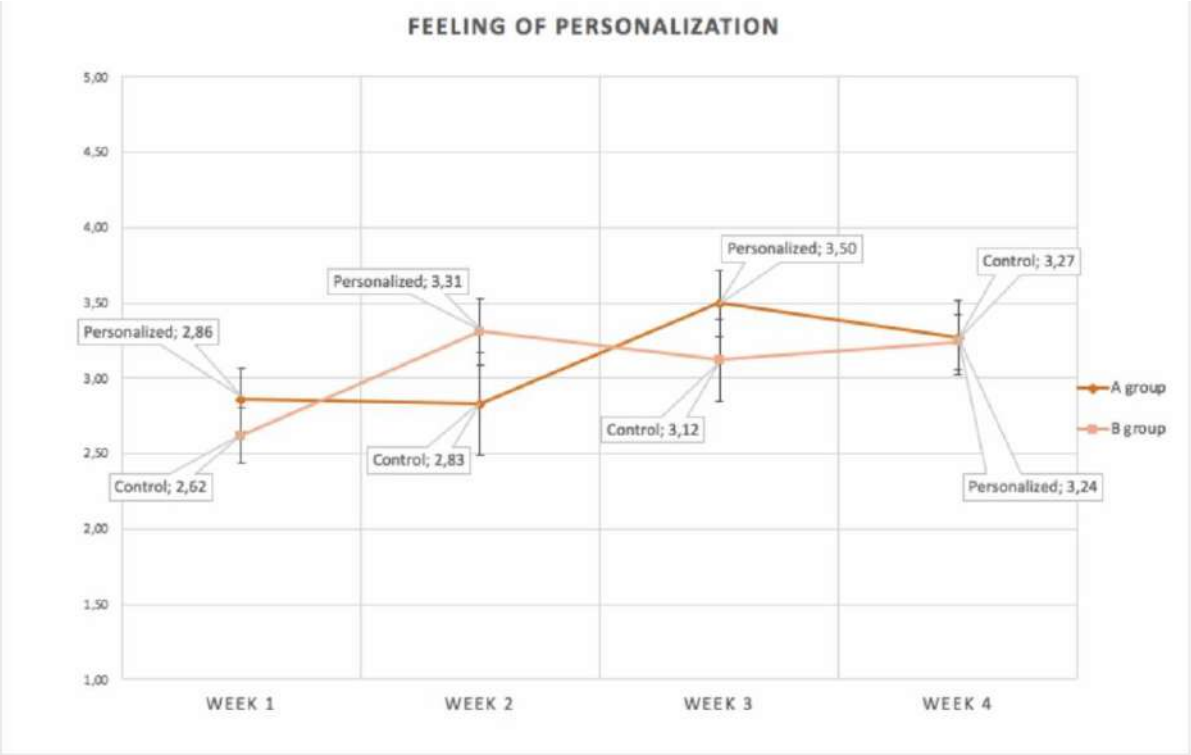


Figure 27: Means of group A and B over the four weeks. Error bars denote the standard error of the mean

➔ There are no group differences between group A and B in any week for feeling of personalization. And if we look at the results within a group and compare their scores per week we find no statistical difference.

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

Question 3: "*I have the feeling I have been missing out on something".*

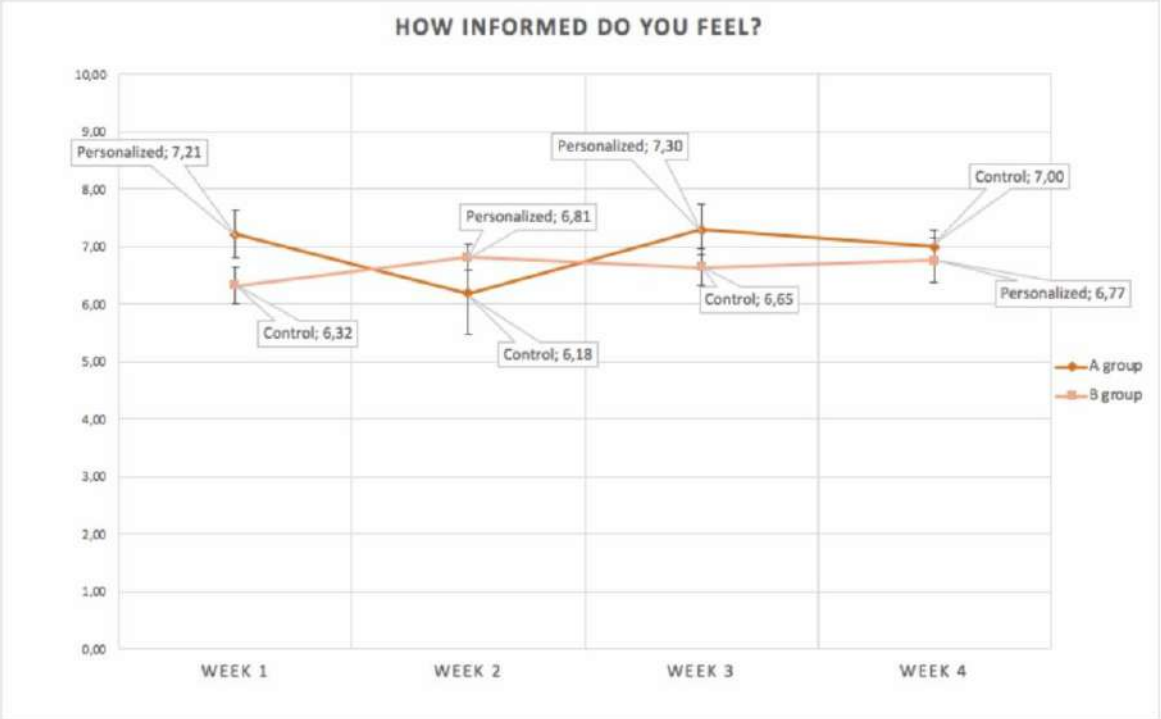


Figure 28: Means of group A and B over the four weeks. Error bars denote the standard error of the mean

There are no group differences between group A and B in any week for Feeling of Missing Out (FOMO). And if we look at the results within a group and compare their scores per week we find no statistical difference.

Combining Relational User (operational) data and survey (experience) data:

Is clicking more news articles or reading more articles in general related with feeling of being informed?

➔ In general, we can conclude that there is a positive relation of number of clicks and feeling of being informed in this pilot over the 4 weeks. (The detailed analysis can be found in the annex.)

Is there a relation between clicking, reading time and scrolled full with feeling of being informed per group and per week?

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

WEEK 1

The only meaningful correlation in Week 1 is in the personalized condition between number of clicks in personalized stream and the feeling of being informed.

WEEK 2

The only meaningful correlation in week 2 is in the control condition between number of clicks in personalized stream and the feeling of being informed. Other are no meaningful correlations between feeling informed and any usage metrics of the personalization stream in week 2. (Detailed analyses can be found in the annex.)

WEEK 3

There are no meaningful correlations between feeling informed and any usage metrics of the personalization stream in week 3 (detailed results are available on request, but for brevity are not fully reported here).

WEEK 4

There are no meaningful correlations between feeling informed and any usage metrics of the personalization stream in week 4 (detailed results are available on request, but for brevity are not fully reported here).

Advanced Analysis: Regression analysis.

What are the most important predictors of feeling of being informed?

We performed a linear regression analysis with the following independent variables: "Number of clicks in each stream" and "condition: personalized vs. control" and the respective interaction between condition and number of clicks in each stream.

➔ The regression analysis shows that these variables, number of clicks per stream and condition, are not significant predictors of feeling of being informed. Therefore, we cannot conclude that in this research number of clicks in a stream or condition predict any changes in feeling of being informed. However, keep in mind that the number of participants is quite small to perform these analyses.

For further details on the calculation please see the annex.

5.3.3 Qualitative Interview Evaluation

Following the four weeks of pilot evaluation, the user partners in Belgium (VRT) and Cyprus (DIAS) reached out to their participants in order to invite them for a live round of feedback in group or individual qualitative interviews. In Belgium, 7 participants were interviewed (1 group interview with 3 people + 4 individual interviews) and in

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

Cyprus, 9 people were interviewed (9 individual interviews). In the following parts, we will discuss different important findings regarding the CPN recommender.

VRT interviewed 7 participants.

DIAS proceeded with personal interviews of the users. 9 users were interviewed. The interviews were conducted via Skype due to holiday season.

DW did not set up a focus group because the participants all live in different time zones and a focus group, even as an online teleconference version would not be feasible. Instead, DW will look into the possibility to set up small local focus groups at a later time. The problem here is that a local focus group would not necessarily represent DW's international target audience.

The following results came out of the discussions:

Feedback on the recommender

Most people we interviewed said to use the app on a daily basis, and even multiple times a day. Checking the app to get a quick news update became a habit throughout the 4-week test period. In general, while using the application, the respondents always felt up to date about current news affairs.

> *"I was already looking for a while for an app with high quality news. I really created a habit of checking the news when I had a couple of minutes of time."*

Positive points

- First, the respondents were asked what they liked about the app. The app was considered easy to use and the interface was intuitive. Participants could obtain Quick update on the current news with a scroll. The mobile app, in comparison to the web recommender in pilot 1, was also seen as a big improvement. There is a big preference for a mobile app as it reminds you to read the news, and enables you to check the news quicker.

- The CPN concept with the 3 streams was evaluated well. The personalized stream worked well, and in combination with the other streams the participants felt they got a complete news update.

- Swiping left/right was easy to indicate interest in an article. This is a very quick and intuitive way to give feedback and is also used in other applications as a navigation tool, which makes it easy to understand for users.

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

Negative points

Next, the respondents were asked what they didn't like about the app.

- In each stream, there was only a limited number of news items. The quantity of news in each stream wasn't considered sufficient. The personalized stream was not frequently renewed and the articles in the most read stream were also outdated. As was also found in the VRT sidetrack, people prefer news not to be older than 2 days as they consider older news irrelevant.
- Interactive elements and videos are not always working and/or not displayed correctly.
- When trying to swipe an article left or right, sometimes you accidentally move to another stream, which interrupts the user experience.
- In the first two test weeks, it wasn't possible to undo a choice you made by mistake (e.g. swiping). (This was fixed in the app update launched halfway the pilot as a reaction to the intermediate user feedback).
- There is only one news source in the application (e.g. VRT NWS for the pilot in Belgium). This was no problem for this pilot test, but in real-life it would be considered a big added value if a multiple news sources could be integrated in the app.
- Some people thought the interface was a bit boring and not attractive.

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

What's missing

The respondents were also asked about features or functionalities they thought were missing. The following elements we mentioned:

- Related content at the end of an article. Some participants liked the option of having related news linked to an article. This way they can follow up on the topic when they are finished reading the article.
- A search function. This would allow users to look for certain keywords or topics.
- A 'read later' button. Now articles disappear from the stream after a certain amount of time. In this case it would be interesting to have a button that allow users to save an article for consumption on a later moment.
- An indication of the categories a news item belongs to subscribe to: some respondents said they tend to seek for categories of news to choose from, just like they do now while they are browsing a website (e.g. politics, local news etc.)
- Mood/moment in the day as context factor: it is important for the app to understand that users are not always in the same mood. For example, during the day they might be more interested in hard news and in the evening, they might enjoy to read articles about culture, lifestyle etc. In addition, during vacations reading habits are altered. How will the app deal with such context-bound changes in interests?
- The option to read the news chronologically.

Overall feedback on personalization

Feedback on the 3 streams

The CPN-concept with the 3 streams was evaluated well. The combination of the 3 streams gives the participants the feeling they get a complete news update. They like to have a personalised stream with a personalized news offer, but they want to have the option to see other feeds in order to tackle their fear of missing out.

> *"It's kind of a hold to also be able to go to headlines or most read news."*
>
> *"I thought it was an added value that you could compare the personalized feed with the headlines picked by the editor and the most read news."*

Most respondents said to start with reading the personalized stream when they open the app, and then proceed to read one other stream or both other streams. One remark that was made by a couple of respondents is that reading three streams can be bit overwhelming. They suggested to only have two streams, one with personalized news and one other one with breaking news or headlines. The personalized feed can then

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

give opportunity to the user to explore the articles of his preference, while at the same time have the chance to switch to the headlines-news.

Opinion on News personalisation in general

The personalized news feed was not perfect yet, but generally evaluated well and definitely moving in the right direction. There is definitely a need for it, as some users expressed that sometimes they do not have enough time to follow current news or search for interesting articles. However, the comibination with other general news stays important.

> *"I merely see it as complementary. I wouldn't only read personalized news, because you might end up in a filtre bubble where you miss certain things."*

Feedback mechanisms

There were some respondents that were not willing giving the app feedback in order to improve the personalization algorithm. They thought it was a waste of time, as the app must find the way to improve the personalization algorithm automatically by exploring their behavior in the app and preferences. However, most participants like the idea of giving the app feedback in some way, as they want to help in making the algorithm more precise and they would like to control a bit how their personalized news feed is created. This are especially willing to do this when they start to use the app. The swiping feature was evaluated very well, as it is very quick and intuitive.

> *"You can't make it easier than swiping."*

> *"I would be willing to do it for a limited period of time. For example the first days or week. But after that I expect the algorithm to do it."*

The uplifting/depressing article classifier

This is considered as a very interesting concept. On some days, or especially in some moments of the day (e.g. in the evening), people like reading positive news. One respondent said that he currently already organized his Instagram account like this, by following a lot of 'inspirational' profiles, because he likes checking it at the end of the day. Two respondents said they would like the have an extra news stream for this, with only positive news for example.

> ***"Sometimes it can be nice, on a day where there only is negative news, to choose to read something positive."***

> ***"An on/off button would be good. Or a seperate stream would be better. You don't have to turn that on or off."***

Feedback on the experimental set-up

From the interviews, it became clear that not all respondents were aware of the experimental design, in which the personalized stream would change from week to

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

week. The quantitative analysis already indicated that there was no real difference in how informed people felt when consulting the different streams, which was confirmed in the interviews.

For the third pilot set-up, we will therefore adjust our research design. As we are moving to an open pilot, the goal is also that users can test the application in their natural setting, meaning that they would be able to use it as they would use any other app, without strict expectations in terms of time and duration of the test. To also be able to test the specific additional questions for pilot 3, we will combine an open test with smaller pilot activities on each of these issues. This will be further determined and detailed in the pilot planning for pilot 3.

5.3.4 Pilot Conclusion

As explained, the second pilot had many limitations, but also some valuable lessons learned for CPN and pilot 3. Certainly, in combination with the sidetracks and deep-dives, there are some interesting findings for the CPN consortium.

The quantitative analysis unfortunately didn't indicate significant differences between both groups in the experimental design. This is probably also due to the low number of users involved in the test.

Related to news personalization, we noticed that people are positive about this idea, but that there still is an important aspect of FOMO. In the interviews, it becomes clear that people like a certain degree of control over their news consumption (for example in selecting categories, a search function etc.). Leaving everything up to an algorithm still seems to be a step too far for some respondents. This is something we will take up in pilot 3 as an additional point of attention.

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

6 HACKATHONS AND COOPERATIONS

As stated before in chapter 3 on the feature deep dives, there are specific topics that will be tackled by external companies already working in the field. These start-ups were chosen during the two hackathon events organized by Digital Catapult and Wan-Ifra as part of the planned exploitation activities.

The first session focused on identifying potential collaborators, first via an open call (publicised widely via consortium members and social media, open to applications from across the world) and then further through the face to face event held at Digital Catapult's offices in London. This Hackathon of ideas brought together industry experts, technology expertise and consortium members to first identify gaps in the current functionality being developed and secondly to identify potential development partners. Attendees from the media side included such well-known names as The Guardian, Sky or Tamedia.

The collaborations formed during this first event led to cooperation between the participating SMEs and media partners in the consortium. These teams used the time between the first and the second event to further develop the ideas they had formed in the first Hackathon to be more defined potential product offerings.

The second Hackathon event then focused on showcasing the art of the possible. During this event, held as part of the Wan-Ifra Global News Media event in Glasgow, the ideas were pitched to an industry audience who selected a winning entry prior to a networking session. The event attracted 50+ media experts from the conference to attend.

Following the Hackathons the consortium is now going to work closely with the SMEs to further develop the proposed functionality. Funding has been made available for the four selected organisations to bring the functionality to a PoC stage. This will allow for some testing of the solutions as part of the final phase of user testing.

Participating Start-Ups

Overall a total of 10 SMEs, Start-Ups and Universities participated in the hackathon events, four of which decided to support CPN through further development of specific tasks. These four Start-Ups and how they are planning to attribute to CPN is described in the following.

U-Hopper

U-Hopper is an R&D intensive deep-tech SME, developing Big Data Analytics solutions based on Machine Learning and Artificial Intelligence technologies. Headquartered in Trento, a tech hub in northern Italy, U-Hopper was born as a spin-off of an international research center (CREATE-NET/FBK). The company focuses on the research, development and deployment of top-quality ICT tailored services for large and mid-sized businesses, and it operates in various verticals (retail, media, logistics, manufacturing, banking).

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

The idea that U-Hopper developed is called Tapoi. It builds profiles based on customer online actions, allowing businesses to provide targeted and tailored services, personalizing user experience and achieving higher conversion. Thanks to its flexible and scalable technology, Tapoi can be customised to meet the needs of businesses operating in different verticals.

By generating a consumer's map of interests, Tapoi helps to overcome the cold start problem allowing even new (CPN) users to be targeted with the right content, at the right time. Based on these exact profiles, Tapoi is also of help in fighting the filter bubble. By means of a customized algorithm, allowing for the appropriate choice of the parameters, the solution ensures that consumers are exposed occasionally to content that is out of their interest sphere, thereby mitigating the echo chamber effect. Using the social login of third-party social media, Tapoi also eases user management. The identity of a user is guaranteed, thereby reducing the cost and time connected to account management processes (e.g. no need to provide two-factor authentication, recovery password procedures). For the user, this means one less password to remember, and one less factor of friction.

Loomi

Loomi.ai is an intelligent assistant platform that seamlessly orchestrates the flow of all information in your life into one ultimate platform, helping users to overcome information overload in our lives. Loomi learns what matters to users so they never miss an important to-do item or breaking news, but can safely ignore all other pieces of content or messages. Loomi.ai was founded in 2017 by Al Ramich and Surjit Bhachu with the core mission of using the intelligent assistant technology to solve the problem of information overload.

For CPN LoomiAssist Ltd will build a News Entity Ontology proof of concept for the purpose of news content personalisation. Loomi shall provide services to support a successful integration and testing of a hosted application accessible via APIs. The ontology will be for English language only for the initial phase of the PoC and will focus on giving (1) API access to get normalised named entities (entity alias matching) as well as (2) API access to get news category for either news article URL or a set of extracted named entities from an article (news categorisation).

The ontology will be built in a way to support the development of additional languages as well as entity "context matching". The APIs for the proof of concept will not be industrialised which will be included in the subsequent production release of the service.

Kensai

Kensai is working on an AI that reads all sentiment and topics around companies, using hyper-scaling cloud AI in order to formulate precise predictions about a topic. The Startup started applying this technology to analyse large companies before moving onto cryptocurrencies, working with funds and banks to provide accurate sentiment analysis and information extraction. The technology is now offered to brands, to understand their products and imagine new applications.

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

For CPN, Kensai will provide a 'Proof of Concept' based on this technology for a set of possible CPN data. Kensai 'BrandBrain' will realise a realtime monitoring of a keyword on Twitter. A sentiment inference on the relevant tweets will provide statistics as to the quantity of 'Positive', 'Neutral' and 'Negative' tweets in relation to the given keyword from the point of monitoring to the present. This will help improve finding more relevant content for CPN users' interests.

Yoop

Yoop is an award-winning software development company based at the University of Nottingham Innovation Park. It has developed natural language processing and recommender systems for news sources, and has experimented with innovative interfaces aimed at capturing audience perceptions of news content.

Yoop Tech Limited will build a proof-of-concept (PoC) software infrastructure for secure and transparent collection, aggregation and storage of user data. The innovation relies on a networked approach to user authentication and a bespoke data security and transparency tool for increased privacy control. The PoC will be built to support the CPN's machine learning technology using privacy by design principles. The proof of concept will be used to validate that the proposed networked approach can provide a better experience at user sign in, help create richer user profile data and that data subjects have more control over their personal data and an increased default privacy.

The proposed infrastructure for this setup will be composed of a single sign on (SSO) interface and API, a user data collector (user profile data, data API and access control system) as well as a data privacy, security and transparency tool (a user profile, permission and data permission interface, data encryption and communication tool, and access log).

Cooperation Process

To ensure a close connection to both the CPN project as well as the media market, each start-up participating in the second hackathon was paired with a media partner (DIAS, DW or VRT) in the project. Some of these pairs already found each other during the first hackathon, others came to be during the first and second event, after all ideas where in the table and a best fit could be found.

DIAS started working with U-Hopper SLR to examine ways to include their service in the CPN project. The first calls where on how to best integrate the service TAPOI in the DIAS system, but DIAS and U-Hopper are already in the discussions of which part will best serve DIAS' users.

DIAS is also in contact with Yoop, as their solution might be of value for DIAS as the only private media outlet in the consortium. An implementation here could also prove useful for CPN as the consortium could directly evaluate the value of the technical solution in a live system and evaluate the complexity of its implementation.

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

Deutsche Welle started interacting with Loomi.ai to address the issue of a proper, standardized media categorization system and keywording of articles. Both keywords and categories are, as mentioned before, a prerequisite to automate processes inside a media company (e.g. creating automated newsletters, suggest related articles o users) and the keystone to deliver good recommendations. With its experience in this field, Loomi is a fitting partner to create a PoC for this, and Deutsche Welle has a natural interest in this, especially with its many languages and different departments.

VRT will be pairing with Kensai, as it has already been working with IMEC on sentiment-related analysis. The outcome is particularly interesting to VRT as it is also taking a closer look into other recommendation possibilities and into including existing social networks.

The plan is to have these PoCs ready at the end of year three, so that they could possibly be included a final product release of CPN at the end of the project.

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

7 CONCLUSION AND NEXT STEPS

In this document, we have described the activities and results from the Pilot 2 development and evaluation phase. We have given insights into the feature deep dives the consortium has undertaken to get a better understanding of some of the elements of personalisation and we have described the sidetracks, VRT and DW have worked on, to prove the interoperability and the flexibility of the CPN recommender-system outside the CPN framework. In the main chapter of the document, the Pilot 2 evaluation of the second version of the CPN prototype was described, from the set-up of the pilot, to the recruitment of the participants to the realization and the Pilot 2 results of the evaluation.

Finally, the document gave an overview of the ongoing works in regard to the hackathons and the collaboration with the startups with which the consortium aims to tackle some further topics to improve the recommendation system. Working on the feature deep dives proved once more that creating the perfect recommendation is a difficult thing.

While these in-depth sessions gave a lot of insights, they also took up more time than expected, both in the setup and evaluation phase as well as in regard to implementation. It will be a challenge in year 3 to filter out the right aspects from these small tests and included others if necessary as well as to see how far they can be integrated into the system and how well they will improve the output. The results from the sidetracks were equally helpful, and equally difficult to realise. While VRTs sidetrack was completed, the consortium hopes to be able to also finish the DW sidetrack in year 3.

Applying the CPN recommender in a native VRT environment proved very useful in order to get more people onboard to test it, but also showed where the difficulties where in regard to implementation of the CPN system outside its own ecosystem. Besides the technical learnings, the testing at VRT also gave us valuable insides into users' thoughts on recommendation. The recency of the articles is of utmost importance in news recommender systems as is the quality of the recommendations. A test with a popularity score (most popular articles) works well, but users don't see it as real personalisation. Overall, this test gave the consortium valuable feedback for further development inside the project and it will be continued (DW sidetrack) and repeated where possible.

Finally running the full Pilot 2 evaluation with all three user partners was the big test for the system giving us more insights into both the technical side as well as the acceptance from the users' side towards a personalised offer. As described in Chapter 5.1, the test was set up with two user groups, comparing behaviour and survey results between users that received personalization and users that did not in the mobile CPN app.

It quickly became clear that the mobile version was much preferred by users over the website version in Pilot 1. But it also clearly showed (e.g. in comparison to the sidetrack

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

at VRT, but also through the feedback) that users have a clear brand loyalty in their media consumption. Some users thought CPN was a new brand and didn't connect it to DIAS, DW or VRT, which had an influence on their use of the system.

This was probably one reason why the user numbers weren't as high as expected, despite the efforts by all user partners to recruit and also to motivate users for and during the evaluation. While this still led to some clicks, it wasn't enough to really get tangible results from the analysis of the numbers or the surveys for DIAS and DW – which makes it harder to detect a clear preference towards or against personalised news. Still the feedback gathered from the users through the surveys and following focus groups and interviews showed that users liked the app in general. Most of the users liked the interface of the app and stated that the app was user friendly, simple and easy to navigate.

Taking the results from the Pilot 2 evaluation into account, there are several things the consortium sees as necessary to change in order to improve the test-results and prove the usefulness and functionality of its recommender-system. While the more obvious changes are in regard to the user interface and the content recommended (like names of the streams, labels on the content and more recent content) there are other aspects concerning the setup and the system itself that need to change.

For one it is necessary to separate the user evaluation from the further evaluation of the recommending system. The consortium needs to be able to check which aspects are actually changing the recommendations how far, to then have users test specific outputs and review whether they are helpful or not in order to be informed. Such a testbed is currently being planned.

Regarding the numbers of participants, the user partners have to increase the advertisement for the plot evaluations. For the pilot 3 evaluation phase, it is planned to change to an open evaluation. This means that recruitment isn't limited to consortium partners. This aspect has already been taken into further consideration by the partners and different strategies are being evaluated, such as professional recruiting outside the consortium, additional, more directed advertisement through social and partner networks.

The aspect of enlarging the possible user group by expanding development into other operating systems (such as iOS) is also part of the discussion and will depend on resources. During the Pilot 2 evaluation it clearly showed that some users would have wanted to participate but couldn't because they didn't have an android phone. Another aspect that will have to be approved is seamless communication across the whole evaluation. Using e-mail to reach users, to lead them to the application for the actual test and then back to a questionnaire on the web via e-mail proved too disruptive. A possible solution could be to integrate the feedback feature directly in the app. This way users wouldn't have to leave the app or fetch their mails to be reminded of giving feedback about what they liked.

The consortium is also discussing to broaden the evaluation approach. While in this case the focus was on getting users to use the app and evaluate their satisfaction and

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

the success of the app mainly via the clicks they made, it might make sense to split the testing. While all users would be analysed based on their clicks again, a smaller group would be directed to evaluate specific features in a more focus-group based setup.

Taking all of these measures into account, the consortium is confident to both reach out to a larger group of testers and offer them a more exciting and hence easier to follow approach to evaluating the CPN application. This will in return lead to more tangible results for the consortium in Pilot 3.

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

APPENDIX A

Detailed statistical survey results VRT Sidetrack

Question 1

Descriptive Phase 1

Sample Size	Median	Average	Confidence Interval of Average	Standard Deviation	Minimum	Maximum
117	7	6.60	6.343 to 6.854	1.40	4	10

Descriptive Phase 2

Sample Size	Median	Average	Confidence Interval of Average	Standard Deviation	Minimum	Maximum
273	7	7.10	6.921 to 7.270	1.46	4	10

Difference between phase 1 and 2

The two-tailed P value equals 0.0018. By conventional criteria, this difference is considered to be statistically significant and hence people that have answered the question after Phase 2 feel on average more informed after reading the tab 'My News' than the people that have answered the question after Phase 1.

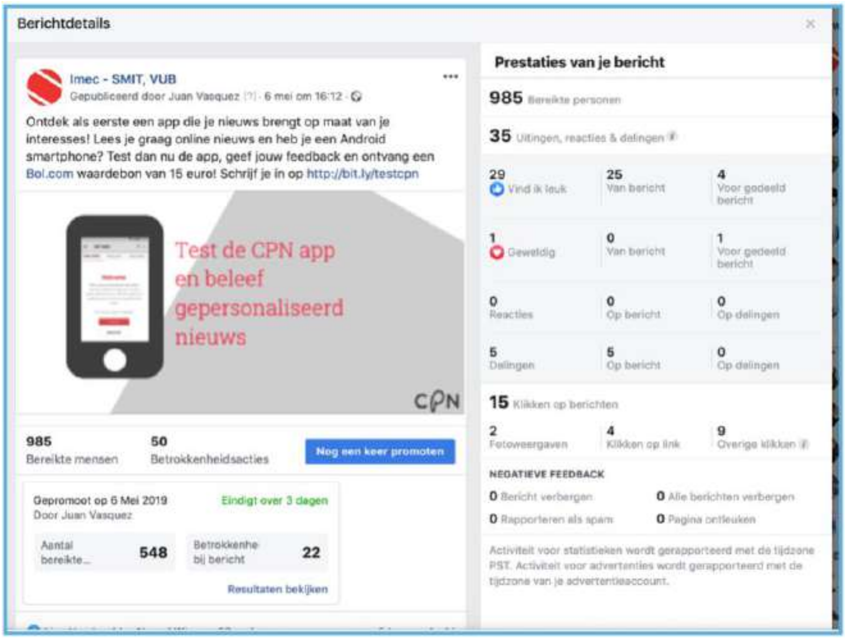
However, keep in mind that this analysis is not exactly methodologically sound, as these groups are not independent from each other. Some people have responded to Survey 1 and Survey 2. Therefore below, the paired results for this group of people will be outlined.

Specifically, we were able to use the IP-addresses to join the responses from the two surveys. We have data from 44 people that responded to both surveys. For these 44 respondents the mean of 'feeling of being informed by reading 'MyNWS' stream' after phase 1 is 5.98 (SD=2.04) and after phase 2 is 6.93 (SD=1.26). Hence, people have a higher feeling of being informed after phase 2 (M difference= -0.95, $t(43)=-3.64$, $p < 0.001$). This seems logical, because in phase 1 they mostly received articles that were old and not updated accordingly. Also note that this a mean over the groups in the respective phase 1 and 2.

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

Recruitment for Pilot 2

Recruitment Article Performance VRT



Recruitment Article Performance Dias

	Page	Pageviews	Unique Pageviews	Avg. Time on Page	Entrances	Bounce Rate	% Exit	Page Value
		1,619 % of Total: 0.00% (88,792,366)	1,513 % of Total: 0.00% (69,469,737)	00:06:18 Avg for View: 00:02:13 (183.38%)	1,378 % of Total: 0.00% (32,853,386)	52.10% Avg for View: 35.88% (45.21%)	72.76% Avg for View: 37.00% (96.65%)	€0.00 % of Total: 0.00% (€0.00)
1.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn	1,594 (98.46%)	1,497 (98.94%)	00:06:13	1,364 (98.98%)	52.56%	73.15%	€0.00 (0.00%)
2.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR3N7Ck2jTmaL_9kVATkQ1sWr1xYoXDo__CVmbkqCNdPoOXkeuFLhzmMflM	8 (0.49%)	2 (0.13%)	00:07:02	0 (0.00%)	0.00%	25.00%	€0.00 (0.00%)
3.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR3IcjdLIuHLcmaH7iSZYrAfFnulMGN635xDASUalZiuYqD0PrxozOjWqk	3 (0.19%)	1 (0.07%)	00:01:26	1 (0.07%)	0.00%	33.33%	€0.00 (0.00%)
4.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR0mEFFp3eZG0QXGLEOLE0I9FCTfq28airRP8TFO1zVuBabloPM06y-Mv14	2 (0.12%)	1 (0.07%)	00:31:47	1 (0.07%)	0.00%	50.00%	€0.00 (0.00%)
5.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR01xviJfsNSIY3LBITPpJeiJrNCLed.Z4NF7HfX8fnoGWMOE03TESN5tSU	1 (0.06%)	1 (0.07%)	00:00:00	1 (0.07%)	0.00%	100.00%	€0.00 (0.00%)
6.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR0i1vwdGhc4aAmFLBnG1dHWzrY0ARstazIR7DRqnvAL8ceNUTFhXoliqDQ	1 (0.06%)	1 (0.07%)	00:00:00	1 (0.07%)	0.00%	100.00%	€0.00 (0.00%)
7.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR0OUnwZlzW5aVi1Lw8VqAgMoacEh1t-HnMoWowYe32e-sJhgbDJ-g-IBys	1 (0.06%)	1 (0.07%)	00:00:00	1 (0.07%)	0.00%	100.00%	€0.00 (0.00%)
8.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR0QTRZEtpb-Byib2VSvQrU_cCnB259eAyWom_MHXcYs6sr9S9ih_MNinM	1 (0.06%)	1 (0.07%)	00:00:00	1 (0.07%)	0.00%	100.00%	€0.00 (0.00%)
9.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR1K6TU2jJZEXanuLJzGndGziyBchDqyZfl29aRMNYTbh4NEWOF2CDyRmAc	1 (0.06%)	1 (0.07%)	00:00:00	1 (0.07%)	0.00%	100.00%	€0.00 (0.00%)
10.	/news/local/561900/thelo-ekseligmeni-enimerosi-cpn?fbclid=IwAR1ZngqPlarvO6zIyAPnV39ziDUrTA3hw0Lr6tfmKK2OVNtUUG8nB1zxaJU	1 (0.06%)	1 (0.07%)	00:09:05	1 (0.07%)	0.00%	0.00%	€0.00 (0.00%)

Rows 1 - 10 of 16

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

Report

5 Apr 2019 to 16 Apr 2019

1 row

Creative	Ad server				
	Ad server impressions	Ad server clicks	Ad server average eCPM	Ad server CTR	Ad server CPM and CPC revenue
700x470	54,190	321	€0.00	0.59%	€0.00
Totals (1 rows)	54,190	321	€0.00	0.59%	€0.00

Dias Recruitment Activities

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

Usage and Survey data combined

Analysis for Phase 2

n = number of participants in the analysis

r = pearson's correlation

p = p-value, by conventional criteria when a p-value is below .05 it is said to be statistically significant.

Overall

A. Correlation between **total articles read** in phase 2 and feeling of being informed n = 113, r = .04, p = .65. There is no significant relation between articles read in phase 2 and feeling of being informed.

B. Correlation between total articles read in phase 2 and feeling of personalized news (Q17_1); n = 110, r = -.01, p = .89. There is no significant relation between articles read in phase 2 and feeling of personalization.
C. Correlation between total articles read in phase 2 and feeling of missing out (Q17_2); n = 108, r = .03, p = .79. There is no significant relation between articles read in phase 2 and feeling of missing out.

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

Clicks Phase 2 per group
Group 1
- Correlation between **clicks in phase 2** and feeling of being informed *n* = 46, *r* = .03, *p* = .85.

Group 2
- Correlation between **clicks in phase 2** and feeling of being informed *n* = 32, *r* = .22, *p* = .23.

Group 3
- Correlation between **clicks in phase 2** and feeling of being informed *n* = 33, *r* = -.17, *p* = .34.

Click Ratio MyNWS Phase 2
Group 1
- Correlation between **click ratio MyNWS stream** and feeling of being informed *n* = 35, *r* = .22, *p* = .21.

Group 2
- Correlation between **click ratio MyNWS stream** and feeling of being informed *n* = 15, *r* = -.09, *p* = .75.

Group 3
- Correlation between **click ratio MyNWS stream** and feeling of being informed *n* = 21, *r* =- .16, *p* = .47.

Analyses from Pilot 2

Between group differences each week.

Only in week 1 there is a slight difference between the personalized and control group, which is borderline significant, $t = -1.79$; df = 34, $p = 0.09$. All other weeks, there is no statistical difference. (t = student's t value, df = degrees of freedom)

Within group differences.

Group A started in the personalized news offer, whereas Group B received randomized articles in their 'Your News' feed. Are there differences in feeling of being informed within the same group from week to week if they go to the other condition?

For Group A, there are no significant differences between weeks. Note that the groups become very small, with only 7 participants to make the comparison between week 1 and 2.

For Group B, there are no significant differences between weeks.

(operational) data and survey (experience) data

Is clicking more news articles or reading more articles *in general* related with feeling of being informed?
- WEEK 1: There is a small, but significant, positive correlation between the *number of clicks* and feeling of being informed in week 1, *r* = 0.28; df = 50, *p* = 0.04.
- WEEK 2: There is a small, borderline significant, positive correlation between the *number of clicks* and feeling of being informed in week 2, *r* = 0.28; df = 42, *p* = 0.06.

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

- WEEK 3: There is a small, but significant, positive correlation between the *number of clicks* and feeling of being informed in week 3, $r = 0.35$; df = 38, $p = 0.03$.
- WEEK 4: There is a small, borderline significant, positive correlation between the *number of clicks* and feeling of being informed in week 4, $r = 0.26$; df = 47, $p = 0.08$.

WEEK 1

Group A Personalized

- Number of clicks in personalized stream and informed: $r = 0.49$; df = 12, $p = 0.07$.
- Read time in personalized stream and informed: : $r = 0.19$; df = 12, $p = 0.52$.
- Scrolled full and informed: : $r = -0.02$; df = 12, $p = 0.94$.

Group B Control

- Number of clicks in personalized stream and informed: $r = 0.01$; df = 20, $p = 0.97$.
- Read time in personalized stream and informed: : $r = 0.00$; df = 20, $p = 0.99$.
- Scrolled full and informed: : $r = 0.31$; df = 20, $p = 0.15$.

The only meaningful correlation in Week 1 is in the personalized condition between number of clicks in personalized stream and the feeling of being informed.

WEEK 2

Group A Control

- Number of clicks in personalized stream and informed: $r = 0.60$; df = 9, $p = 0.05$.
- Read time in personalized stream and informed: : $r = -0.10$; df = 7, $p = 0.8$.
- Scrolled full in personalized stream and informed: : $r = 0.46$; df = 7, $p = 0.22$.

Group B Personalized

- Number of clicks in personalized stream and informed: $r = 0.01$; df = 20, $p = 0.97$.
- Read time in personalized stream and informed: : $r = 0.00$; df = 20, $p = 0.99$.
- Scrolled full and informed: : $r = 0.31$; df = 20, $p = 0.15$.

Detailed operational results VRT Sidetrack

Overall usage statistics

The figure below shows the daily clicks throughout the test period. We notice the following:

- A total number of 949 users registered in the app.
- The number of daily active users decays
- There are temporary peaks after the motivation emails

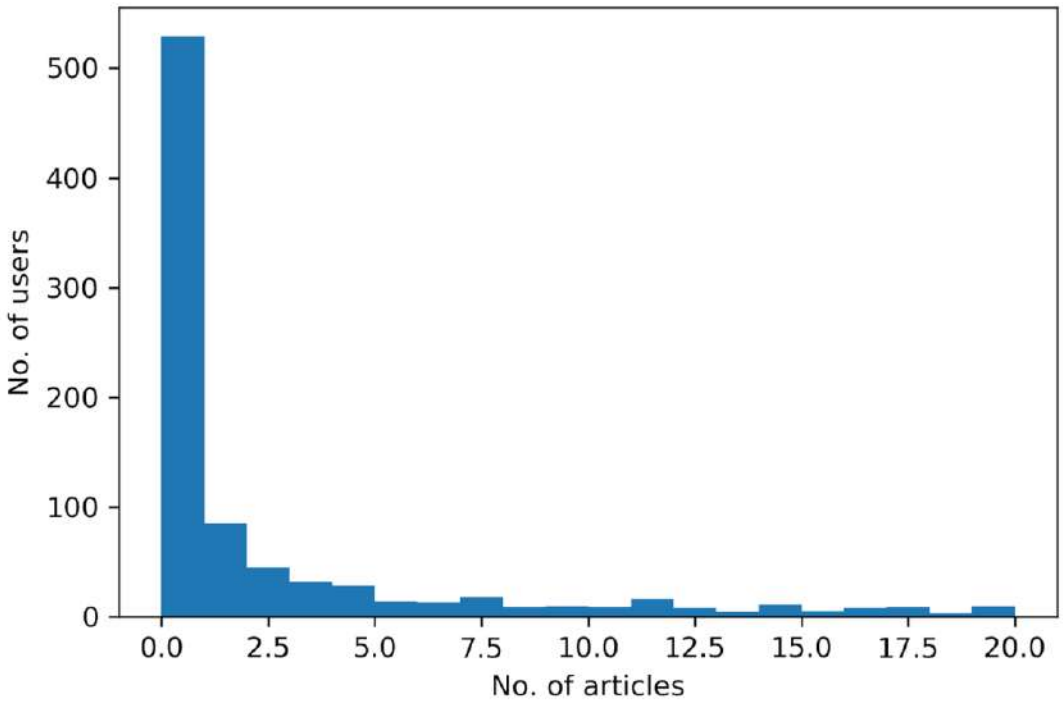
Co-funded by the Horizon 2020 Framework Programme of the European Union

Phase 1

Unfortunately, the technical problems in phase 1 made in-depth analysis of user behaviour infeasible.

Total clicks

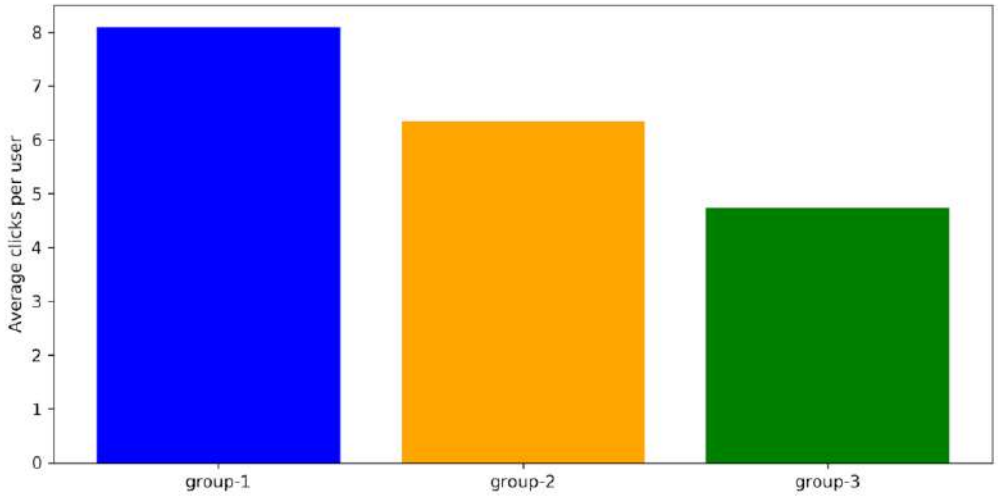
The test users read on average 6 articles during phase 2. The figure below shows a histogram of the clicks. This distribution is leaning to the left: 529 users have read no articles anymore during this period. The average is high because of a relatively low number of heavy users. The highest number of articles read by one user during this period is 150 (not visible in this trimmed histogram).



Clicks per group

We observed a difference in the average number of clicks per user between the three test groups, see figure below. Group 1, the group with the articles that were trending during the last hour, has on average the highest number of clicks per user.

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



All groups contain a small number of heavy users, which have a big influence on the averages. These outliers are bad news for statistical relevance.

We performed the Wilcoxon rank-sum test to check if the differences between groups are relevant.

Null hypothesis	p-value
group-1 = group-2	0.025
group-1 = group-3	0.072
group-2 = group-3	0.5160

The low p-value for null hypothesis group-1 = group-2 allows us to reject it.

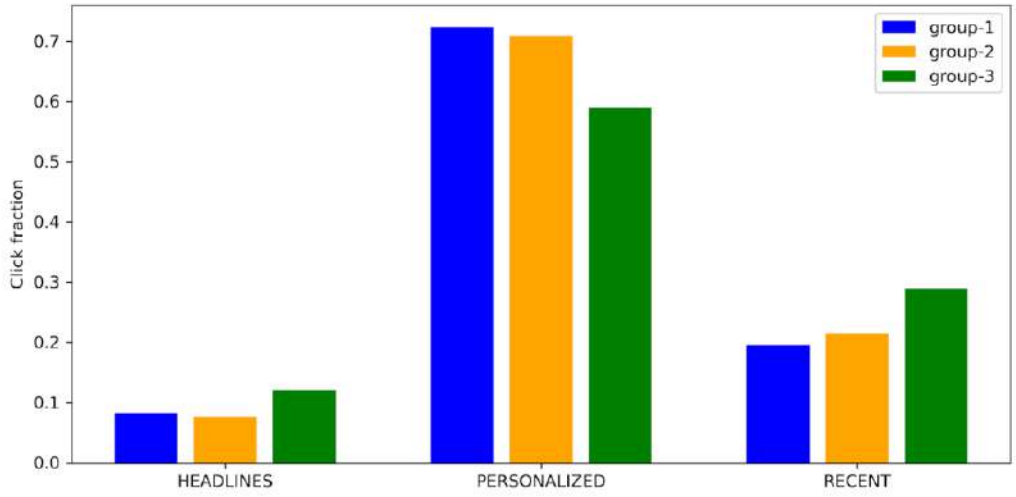
Conclusion:

Calculating POP on the last one hour of data leads to more clicks than on the last twelve hours.

Click ratios per stream

We aggregated the clicks per user, and then calculated the ratios of clicks in each stream. The figure below shows the mean click ratio for each stream per group. For all groups, the headlines stream is very unpopular. Users in group-1 and group-2 prefer the personalized stream more than users in group-3. Users in group-3 have more interest in the RECENT stream than others. This is probably a consequence of the fact that the PERSONALIZED stream for this group contains older items.

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



We performed the Wilcoxon rank-sums test on the click ratios in the personalized tab.

Null hypothesis	p-value
group-1 = group-2	0.961
group-1 = group-3	0.010
group-2 = group-3	0.021

Conclusion:

Calculating POP on the last two days results in lower average click ratio in the personalized news stream than calculating POP on the last hour or the last 12 hours.

Comparison between phase 1 and phase 2

The table below shows the large differences between phase 1 and 2. The number of active users, i.e. users that had at least one click, dropped to almost half. The number of clicks decreased with a factor of four. However, the ratio of clicks in the personalized stream was much higher during phase 2 than during phase 1.

	Phase 1	Phase 2
Active users	835	420
Total clicks	22274	5415
Click ratio in MIJN NIEUWS	0.38	0.68

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