

Grant Agreement No.:



761488



D5.4: Final report on dissemination



Work package	WP 5
Task	T5.3
Due date	April 30, 2020
Submission date	April 30, 2020
Deliverable lead	IMEC
Version	v1.0.
Authors	Wendy Van den Broeck (imec), Stephen Fozard (WAN-IFRA), Teemu Henriksson (WAN-IFRA), Natasja Van Buggenhout (imec)
Reviewers	Olga Kisselmann (DW), Tilman Wagner (DW)
Keywords	Dissemination, communication, dissemination material, events, website, social media, publications

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	15.04.2020	Table of contents	Wendy Van den Broeck
V0.2	22.04.2020	Writing of deliverable	Wendy Van den Broeck (imec), Stephen Fozard (WAN-IFRA), Teemu Henriksson (WAN-IFRA), Natasja Van Buggenhout (imec),
V0.3	24.04.2020	Internal review	Olga Kisselmann (DW), Tilman Wagner (DW)
V0.4	29.04.2020	Completing document	Wendy Van den Broeck (imec) Stephen Fozard (WAN-IFRA), Teemu Henriksson (WAN-IFRA)



DISCLAIMER

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

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

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



EXECUTIVE SUMMARY

This deliverable is the final report on dissemination and communication activities from the CPN project following-up on deliverables D5.1, D5.2. and D5.3. It describes the communication and dissemination actions that took place in the final stage of the CPN project from October 2019 to April 2020, together with the communication linked to the pilot 3 activities.

The communication and dissemination activities within CPN had multiple aims: providing information about CPN goals and progress towards a broader community, the presentation of CPN activities during different events to different target audiences (e.g. industry, academics), and the engagement of relevant stakeholders (e.g. media companies, SME's). This final deliverable provides an overview of these final communication and dissemination efforts and highlights their outcome.

The exploitation activities and results of the business clinics are reported in D5.5.



TABLE OF CONTENTS

1	Final report on Dissemination and Communication activities	9
1.1	CPN website	13
1.1.1	Changes and additions to the website	13
1.1.2	Website traffic	16
1.1.3	Promotional material on the website	20
1.1.4	Pilot Section	20
1.2	Blog posts and Newsletter	22
1.3	Social media	25
1.3.1	Twitter account	25
1.3.2	Instagram account	26
1.3.3	Facebook account	26
1.3.4	Social media campaign for pilot 3	27
1.4	Publications	30
1.4.1	Scientific publications	30
1.4.2	Partners' publications	31
1.4.3	Media mentions	32
1.5	Events	33
1.5.1	International conferences	33
1.5.2	Workshops	36
1.5.3	Hackathons	36
1.5.4	Final CPN event (Webinar)	37
1.6	Outreach to external media companies	40
1.6.1	Outreach activities	40
	Call for Interest (CFI)	40
	CPN Information Webinar 8 November 2019	42
	Distribution of a "How To Media Company Integration" document	42
1.6.2	Status of External media companies engagement (see D5.5)	43
	Conclusions	44
	Annexes	45



LIST OF FIGURES

Figure 1 Website visitors September 2017-April 2020	18
Figure 2: Screenshots of the CPN Instagram landing page	26
Figure 3: Screenshot of the CPN Facebook page	27
Figure 4: Screenshot of a boosted Facebook post	28
Figure 5: Sponsored post on the Instagram feed and explore page	29
Figure 6: Instagram story advertisement	30
Figure 7 Digital Media Europe 2020 postponed	37
Figure 8 Webinar promotion on Twitter	39
Figure 9: Call for interest CPN website	41
Figure 10 Call for interest CPN Twitter	41
Figure 11: How to Media Company Integration document	42



LIST OF TABLES

Table 1	Status of communication activities- end of project	10
Table 2	Overview of changes to the project website	13
Table 3	Overview of website visits,unique visitors and pageviews per month (data from Squarespace analytics)	16
Table 4	Overview of promotional material on the website	20
Table 5	Overview published blog posts and newsletters	22
Table 6	Total pageviews of blog posts throughout the project	24
Table 7	Evolution of our Twitter account	25
Table 8:	Overview of the results of the boosted Facebook post	29
Table 9	Events organised and attended during the final year of the project	33
Table 10	Events organised and attended during the second year	34
Table 11	Events organised and attended during the first year	35
Table 12	Events planned during the final year of project, but cancelled or postponed due to Covid-19 travel restrictions	36



ABBREVIATIONS

CPN	Content personalisation network
DoW	Description of Work
KPIs	Key Performance Indicators
GAMI	Global Alliance for Media Innovation
WP	Work Package



1 FINAL REPORT ON DISSEMINATION AND COMMUNICATION ACTIVITIES

The detailed dissemination and communication strategy was outlined in D5.1. Communication and dissemination plan. The aim of that plan was “to guarantee wide visibility of the CPN results to relevant stakeholders all around Europe in order to attract a critical mass of potential customers” (CPN Grant Agreement, part B, p. 7).

In this deliverable D5.4, we report on specific dissemination and communication activities and its outcomes that have taken place between September 2019 and April 2020, the official end of the project. Our activities were aimed at different target groups: news media companies, journalists, end-users and academics.

Table 1 below provides an overview of the KPIs presented in D5.1 and in the DoW and a final status on each of these targets at the end of the project. Targets in green have been achieved or exceeded, targets indicated in orange have not been reached completely. As can be seen, only the n° of visitors and visits to the CPN website itself did not reach the foreseen target, all other KPIs were met.



Table 1 Status of communication activities- end of project

Activity	Expected outcome	Final status April 2020)
N° of visitors and visits to the website	20.000 by end of the project (this KPI is interpreted as 'visits' and thus not 'unique visitors')	Total n° of visits, end of project: 8,967 visits to CPN website. (+ 14.111 visits to CPN content on our partner sites)
N° of Twitter followers	300 by end of the project	312 at the end of the project
N° of interactions with blog posts	1000/year	1st year of project: 311 2nd year of project: 1793 3rd year of project: 1104 * * Not a full year: September 2019–April 2020
N° of international conferences attended with specific activities for CPN	Min. 3 per year	1 st year of project: 9 events between September 2017 and August 2018 2 nd year of project: 9 events between September 2018 and August 2019 3 rd year of project: 5 events between September 2019 and April 2020
Hackathons	Min. 2 throughout the project, over 50 attendees, more than 6 additional services (see KPI 6.3 in DoW)	First event organised in February 2019. (attended by 12 external media professionals and 20 innovative technology experts) Second event in June 2019, as part of the World News Media Congress. (About 60 media executives, editors and other professionals in attendance.) 4 startups working on new features for CPN, with 8 additional services piloted through the SMEs partners.

Workshops	Min. 3 throughout the project	<p>1st year of project: 6 workshops with end-users and professional users (n=49)</p> <p>2nd year of project: 6 workshops with end-users (n=31)</p> <p>3rd year of project: no additional workshops but user evaluation activities via open pilot</p>
Project events to engage the community of media companies and technology providers	2 throughout the project (see KPI 6.4 in DoW)	<p>The February 2019 Hackathon was our first Major project only event gathering 12 external media professionals and 20 innovative technology experts in attendance.</p> <p>The original plan was to end the project with a final project event hosted at WAN-IFRA's Digital Media Europe Conference in Vienna on 1 April 2020. The idea was to host a 3 hour workshop including a final demo of the CPN tool and a selection of speakers on the broader topic of personalisation in the news industry, best cases, challenges, discussions on Privacy and transparency. Due to the Covid-19 sanitary crisis, the whole event was cancelled and the team developed a similar remote version of this event through a 90 minute webinar to be held on 7 May 2020 (more details in 1.6.1 Outreach).</p>



Recommendations for the platform	n° of stakeholders that register to participate in the community: 100 (see KPI 6.1 in DoW)	<p>At the end of the project, we consider that the CPN platform is open and has been recommended to more than 100 potential customers (media companies) through several indicators:</p> <ul style="list-style-type: none"> - 20 active expressions of interest from media publishers (see D5.5) - at least 80 sign ups for the final CPN webinar (60+ registered April on 30th) <p>And of course:</p> <ul style="list-style-type: none"> - 3 mailings to 3300 media professionals each from the wan- ifra database - 103 newsletter subscribers - 312 Twitter followers
Publications, presentations and demonstrations	Min. 20 throughout the project (see KPI 6.2 in DOA)	<p>25 blog posts 4 newsletters 19 event presentations</p>

In the following sections, the different channels are described and discussed in detail, namely: the CPN website and its content plan (section 1.1), the blog posts and newsletter (section 1.2), the social media channels and activities (section 1.3), publications (section 1.4) and international conferences, workshops and summits (section 1.5).



1.1 CPN WEBSITE

The official website of the CPN project is <https://www.projectcpn.eu/>.

D5.1 gave an overview of the website content plan, namely its categories and content that would be provided on the website. In this deliverable we discuss the final overview of the website.

1.1.1 Changes and additions to the website

The table below describes how the website sections have evolved over the course of the project.

For the third and final pilot, a separate pilot section was created, explaining what the pilot was about and how people could participate. In the 'Promotion' section, additional pilot information for media companies was also provided. The project dissemination was extended with additional videos, posters and postcards.

Table 2 Overview of changes to the project website

	Initial division (D5.1)	Changes/additions made between D5.1 and D5.3	Changes/additions made between D5.3 and D5.4
News	"This category will contain all news about the project, including activities, presentations at events, newsletters and blog posts containing updates on work packages and project progress by partners, specific challenges within tasks, interesting links and articles on current items related to the project's topic, etc."	<p>The news category still contains news about the project, mainly in an overview of the blog posts and newsletters that were published.</p> <p>The CPN twitter feed was added at the bottom of the page.</p>	On the News page, we added the blog posts and newsletters published since D5.3., leading to a total of 23 blog posts and 3 newsletters.
About the project	<i>Vision:</i> description of the project's vision and aims	<p><i>Vision:</i> description of the project's vision and aims</p> <p><i>Project video</i></p>	No changes.

	<i>Partners:</i> logos and links of all project partners	<i>Partners:</i> logos and links of all project partners	The call to action for external media organisations is still in the partner part. Here news organisations can find all information related to the pilot and how to become a CPN pilot partner.
		<i>Call to action</i> for external media organisations to become <u>pilot partners</u>	
	<i>Work packages:</i> description of the work packages	<i>Work packages:</i> description of the work packages, including a description of the work packages, summaries of the published deliverables related to these work packages, as well as links to the full deliverables	Update of the deliverables.



Publications	"This section will include journal articles, conference publications, event publications, as well as project deliverables and press releases"	<p>Instead of only using the tab 'publications', a new one was made, namely 'Project outcomes' with subsections for work packages and deliverables.</p> <p>In the separate section "Promotion", dissemination material was added.</p>	The project outcomes were updated.
Promotion	/	Project dissemination: This page includes all promotion material about the CPN project that website users can share further: videos, poster, postcards and flyer	Additional videos were created.
	/	<i>Pilot promotion:</i> This section includes promotion material calling for media companies to partner with CPN: social media banners and a flyer.	Update of promotion material for pilot 3.
Events	/	Events: new tab for upcoming events	Update with events.
Contact page	"Provides contact information and a link for a subscription to the newsletter."	Provides contact information and a link for a subscription to the newsletter.	No changes.
		Twitter feed with the latest tweets was added on the contact page.	



Twitter feed	/	Twitter feed is added on the News tab and can be found at the bottom of the page	No changes.
Pilot	/	/	This new section PILOT was created to inform people (general news consumers) about the pilot and how to participate. This page contains more information on the pilot, links to install the CPN app and a list of FAQ.

1.1.2 Website traffic

The table below gives an overview of the number of visits for the website per month, from September 2018 to April 2020:

Table 3 Overview of website visits, unique visitors and pageviews per month (data from Squarespace analytics)

Month	Visits	Unique visitors	Pageviews
September 2017	7	7	25
October 2017	11	7	21
November 2017	54	31	210
December 2017	42	29	108
January 2018	210	198	640
February 2018	380	366	707
March 2018	217	190	459
April 2018	102	114	179
May 2018	144	137	273
June 2018	172	168	367
July 2018	259	231	572
August 2018	184	186	477
September 2018	284	279	616
October 2018	351	314	795



November 2018	270	269	561
December 2018	168	161	375
January 2019	304	235	867
February 2019	364	343	781
March 2019	429	476	883
April 2019	301	291	945
May 2019	625	645	1647
June 2019	320	227	899
July 2019	348	340	1048
August 2019	244	245	596
September 2019	260	244	653
October 2019	249	202	629
November 2019	207	214	588
December 2019	175	148	460
January 2020	344	301	908
February 2020	1567	1594	2239
March 2020	452	453	661
April 2020	296	305	639

Note: the Squarespace analytics defines the data as such:

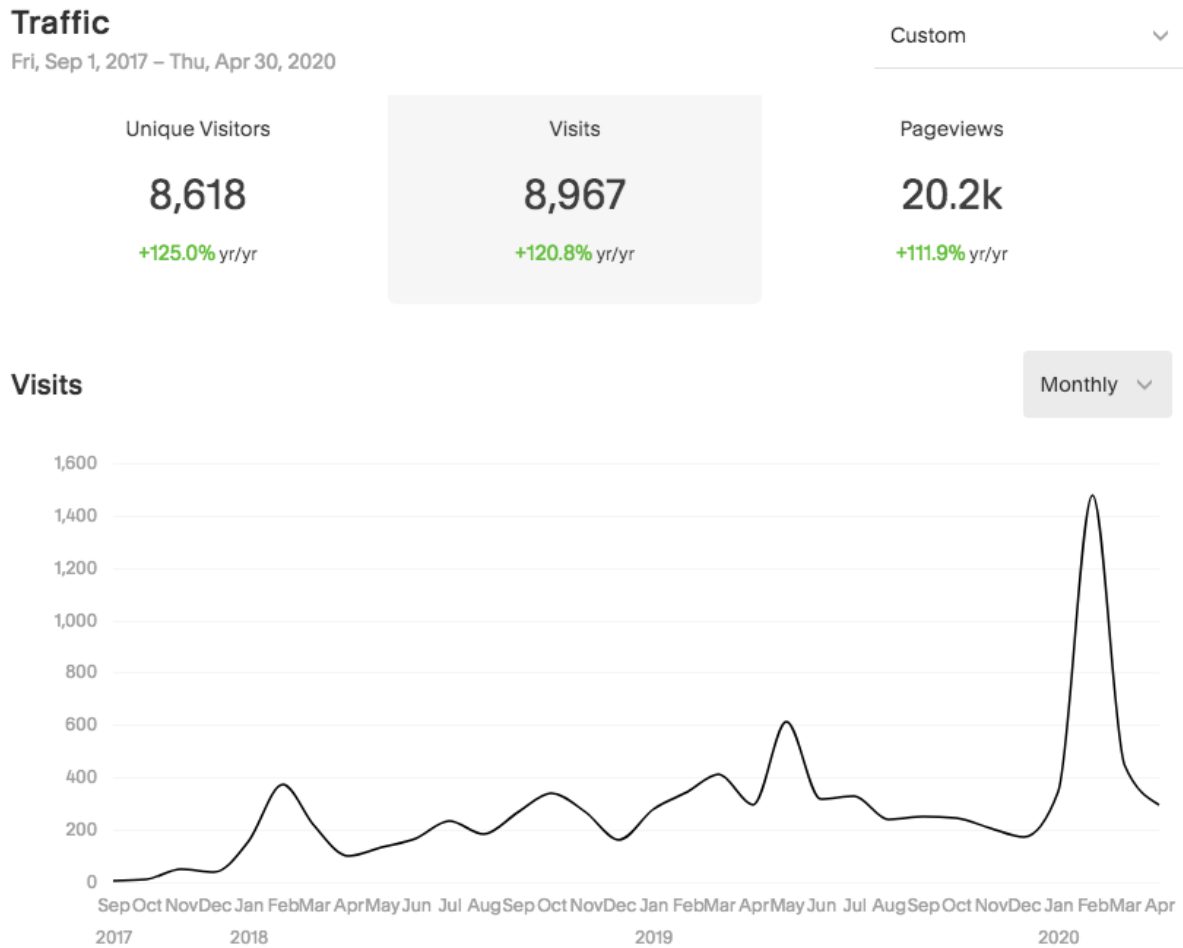
- A Visit is a single browsing session by the same visitor.
- Unique Visitors (previously "Audience Size") is an estimate of the total number of visitors that reached your site in the selected time period.
- A Pageview represents an actual page request the site saw in the time period.

Sometimes the figure for "Unique Visitors" is higher than the corresponding figure for "Visit". This is counterintuitive, but seems to be an issue with the service's monitoring system.

From the start of the CPN project in September 2017 until April 21, 2020, the CPN website reached a total amount of **8,967 visits, 8,618 unique visitors** and **20,2k pageviews** :



Figure 1 Website visitors September 2017-April 2020



The KPI for the number of visits to the website was 20.000 by the end of the project, as mentioned in Table 1. We are behind this goal, despite a sharp increase of traffic in the beginning of 2020, thanks to the promotion activities related to Pilot 3.

However, if we also consider external websites related to CPN, these contribute significantly to the overall traffic numbers.

Related to the **hackathon events**, we created web pages on the Digital Catapult and Eventbrite websites, with the following statistics that were not included in the total above:

- Open Call webpage: Unique views - 467, Total views - **524**
- Eventbrite: **350** pageviews

For **pilot 3**, project partners also shared content on their own websites and social media channels, leading to the following overview:

- VRT MyNWS page on Pilotzone website: **2707 total views**.
- DIAS published 3 articles on Sigmalive Website related to CPN Pilot: The first article¹, informing about the beginning of the pilot, had **9615 total views**. The

¹<https://www.sigmalive.com/news/local/607900/i-nea-apokleistiki-texnologia-tou-sigmalive>



second article² with all the important info regarding the operation of Pilot had **283 views**. Last, an article³ following the interview of the General Manager of Sigmalive, Christos Danezis, who was informing the audience regarding Pilot 3 at Sigma tv's morning news broadcast with **98 views** in total.

- DW used its research related social media channels to promote the pilot and reach out to its followers. The postings on twitter reached about **3500 views** in total. The update on LinkedIn got around **500 views**.

²<https://www.sigmalive.com/news/local/611378/ola-osa-xreizetai-na-gnorizete-gia-tin-xrisi-tis-neas-mas-texnologias>

³<https://www.sigmalive.com/news/local/613222/to-sigmalive-kainotomei-me-ti-dokimastiki-ekdosi-tis-neas-tou-texnologias>



1.1.3 Promotional material on the website

The “Project dissemination” page includes all the promotional materials regarding the CPN project.

In the final period of the project, we focused on the creation of an additional video. The video was set up to summarize the work we did during the CPN project, present the involved project partners, describe how we approached the pilots and present our biggest learnings. We consider the video the ideal tool to communicate in a short and easily understandable way how CPN created value towards external (media) companies that are interested in the matter. We shared the video together with the invitation for the final event webinar, so that industry enthusiasts would be triggered to join the webinar and check out the project results.

Other promotional materials available on the website are as follows:

Table 4 Overview of promotional material on the website

Category	Number	Content
Video	4	goals, hackathons, final working of CPN
Poster	1	objectives, timeline and recommender functionalities
postcards	2	CPN objectives
Flyer	3	CPN and its objectives Call for participation of media companies CPN product board flyer
Social Media banners	3	call for participation of media companies on Facebook, Instagram, Twitter

1.1.4. Pilot Section

For the final pilot, a separate pilot section was created on the website.

This section contains links to the different pilot content offers (DW, DIAS, VRT for testing in English, Greek and Dutch), an FAQ and contact information.

Each project partner also linked to the pilot via their own webpages and social media:

- As mentioned before, **DW** used its strong social media presence to promote the third pilot as the open CPN pilot. It posted several updates about it on the @DW_Innovation Twitter Channel as well as on its LinkedIn Profile always referring

users to the pilot section on the CPN website for further details. The open pilot was also promoted through the general project channels.

- For **VRT** the pilot was mentioned on the Pilotzone webpage & a mailing was sent out to pilotzone testers, a message was posted on the VRT Innovation facebook page and VRT Innovation website, a message appeared on internal screens to inform/attract VRT employees, an article was published on the VRT NWS website, the pilot was mentioned on the VRT website and in several internal & external newsletters.
- **DIAS** published three articles on the website related to pilot 3 and there was a TV-interview broadcasted on morning TV news informing the audience about the CPN goals and the pilot.



1.2 BLOG POSTS AND NEWSLETTER

In the dissemination and communication plan (D5.1), 20 blog posts and 4 CPN newsletters were foreseen throughout the project. We provided regular updates throughout the project and reached a total number of 24 blog posts that were created and published. Besides the blog posts, we also sent out 3 newsletters via Mailchimp. A fourth and final newsletter will be sent out at the end of April, and will include the invitation for the Webinar and the discussion of the final conclusions of the project. Currently, 103 people have subscribed to the newsletter.

The table below shows an overview of the blog posts and newsletters published since the start of the project. The sections highlighted in blue contain the new publications since the previous update on the dissemination activities.

Table 5 Overview published blog posts and newsletters

Blog post number	Content	Title	Published
Blog post 1		WAN-IFRA one of nine CPN partners	23 September 2017
Blog post 2		"We tell you what you want, what you really really want"	27 November 2017
Blog post 3	-	3... 2... 1... Project CPN has taken off!	1 February 2018
Blog post 4	D2.1	CPN: towards the creation of an Open Virtual Platform	2 March 2018
Blog post 5	D1.3	Beyond the state-of-the-art: innovative components in CPN	31 May 2018
Blog post 6	D1.1	(Re-)building trust in personalisation	12 June 2018
Blog post 7	-	Article about personalised news by Titus Plattner	25 June 2018
Blog post 8	Event	CPN @ Deutsche Welle Global Media Forum	25 June 2018
Newsletter 1	Newsletter 1 ⁴	Our newsletter is out	25 September 2018
Blog post 9		The 5 personalisation challenges that CPN is trying to resolve	October 18, 2019
Blog post 10	D4.2	The first CPN pilot is finished!	29 January 2019

⁴ <https://mailchi.mp/f78c41a9b9eb/towards-a-different-approach-for-personalising-news>



Blog post number	Content	Title	Published
Blog post 11	Event	Hackathon 1 report	25 February 2019
Call for interest	-	Calling media companies to pilot CPN	23 April 2019
Newsletter 2	Newsletter 2 ⁵	CFI for pilot 2, Results pilot phase 1, Hackathon report	7 May 2019
Call for interest	-	Calling for test users for pilot 2	7 May 2019
Blog post 12	D3.3	Discover the “technology bricks” that power the CPN platform	14 May 2019
Blog post 13	Report on VRT sidetrack	What we learned from VRT MyNWS	27 May 2019
Blog post 14	Event	Hackathon 2 report	6 June 2019
Blog post 15	-	Highlighting the startups working with CPN	18 June 2019
Blog post 16	D2.3	Introduction: CPN Open Virtual Platform	9 July 2019
Blog post 17	D1.6	How CPN puts users in control of their personal data	19 August 2019
Blog post 18	-	Holding platforms accountable: fines alone won't do. What does this mean for CPN and other EU platforms?	28 August 2019
Blog post 19	D4.3	Second evaluation round – what did we do and what did we learn	November 14, 2019
Newsletter 3	Newsletter 3 ⁶	The next step in personalising news content	18 November, 2019
Blog post 20	Event	Distribution framework: licensing for the 21st century	December 11, 2019
Blog post 21	D3.4	Discover the final versions of the CPN technology bricks	January 21, 2020
Blog post 22		CPN pilots, what we've learned so far and how you can join the next round of user testing	January 29, 2020
Blog post 23		CPN news recommender: how our personalisation solution works	March 12, 2020
Blog post 24		Are filter bubbles really to blame for social and political polarisation?	April 20, 2020
Blog post 25	D4.4	Pilot 3: what we learned from the last round of user testing	April 27, 2020
Call for interest		Join our webinar on news personalisation!	April 28, 2020

⁵ <https://mailchi.mp/e9bc7a681447/towards-a-different-approach-for-personalising-news-635917>

⁶ <https://mailchi.mp/255e822e6350/towards-a-different-approach-for-personalising-news-733825>



One of the KPIs indicates that we aimed to reach 1.000 interactions with blog posts per year. 'Interactions' are often defined as the amount of likes, comments and shares of the posts on a website. We measured **3208 pageviews** across all blog posts, which we consider interactions with the content, throughout the project. We did not receive any direct comments on the articles on the blog. Likes and shares, which could also be counted as interactions, were not available on our blog system.

Since the start of the project, the blog posts received the following pageviews:

- 1st year (September 2017–August 2018): **311**
- 2nd year (September 2018–August 2019): **1793**
- 3rd year (September 2019–April 2020): **1104 (not a full year)**
- The table below includes the pageview breakdown by blog post since the beginning of the project.

Table 6 Total pageviews of blog posts throughout the project

Blog post	Pageviews
3... 2... 1... Project CPN has taken off!	264
From idea to concept: report from the CPN hackathon	258
The first CPN pilot is finished!	229
Highlighting the startups working with CPN	222
The 5 personalisation challenges that CPN is trying to resolve	201
Article about personalised news by Titus Plattner	196
Introduction: CPN Open Virtual Platform	165
CPN: Towards the creation of an Open Virtual Platform	165
Rebuilding audience's trust in algorithms: personalisation at the World News Media Congress 2019	160
(Re-)Building trust in Personalisation	137
What we learned from VRT MyNWS: a pilot on news personalisation at the Flemish public broadcaster	135
Test personalised news content with the CPN app!	115
Discover the final versions of the CPN technology bricks	106
Second evaluation round – what did we do and what did we learn	106
Beyond the state-of-the-art: innovative components in CPN	104
Holding platforms accountable: fines alone won't do. What does this mean for CPN and other EU platforms?	91
How CPN puts users in control of their personal data	83



Discover the “technology bricks” that power the CPN platform	78
CPN pilots, what we’ve learned so far and how you can join the next round of user testing	72
Become a CPN pilot partner and join the next phase of news personalisation	65
Distribution framework: licensing for the 21 st century	60
WAN-IFRA one of nine CPN partners	54
CPN @ Deutsche Welle Global Media Forum	52
"We tell you what you want, what you really really want"	37
CPN news recommender: how our personalisation solution works	29
Are filter bubbles really to blame for social and political polarisation?	24
TOTAL	3208

1.3 SOCIAL MEDIA

1.3.1. Twitter account

The CPN Twitter account, @project_CPN, was used to post regular updates about the project activities, pictures of attended events, announcements of new blog posts and newsletters, and calls for participation in the pilots. In addition, we posted interesting articles, related to personalisation.

The table below shows the evolution of both the number of tweets and the number of followers throughout the project’s lifecycle.

Table 7 Evolution of our Twitter account

		October 2018	March 2019	August 2019	April 2020
number of tweets	of	<i>107 tweets</i>	<i>148 tweets</i>	<i>214 tweets</i>	<i>324 tweets</i>
number of followers	of	<i>147 followers</i>	<i>230 followers</i>	<i>262 followers</i>	<i>312 followers</i>

The KPI for the Twitter account was to reach a minimum of 300 followers, which has been reached.

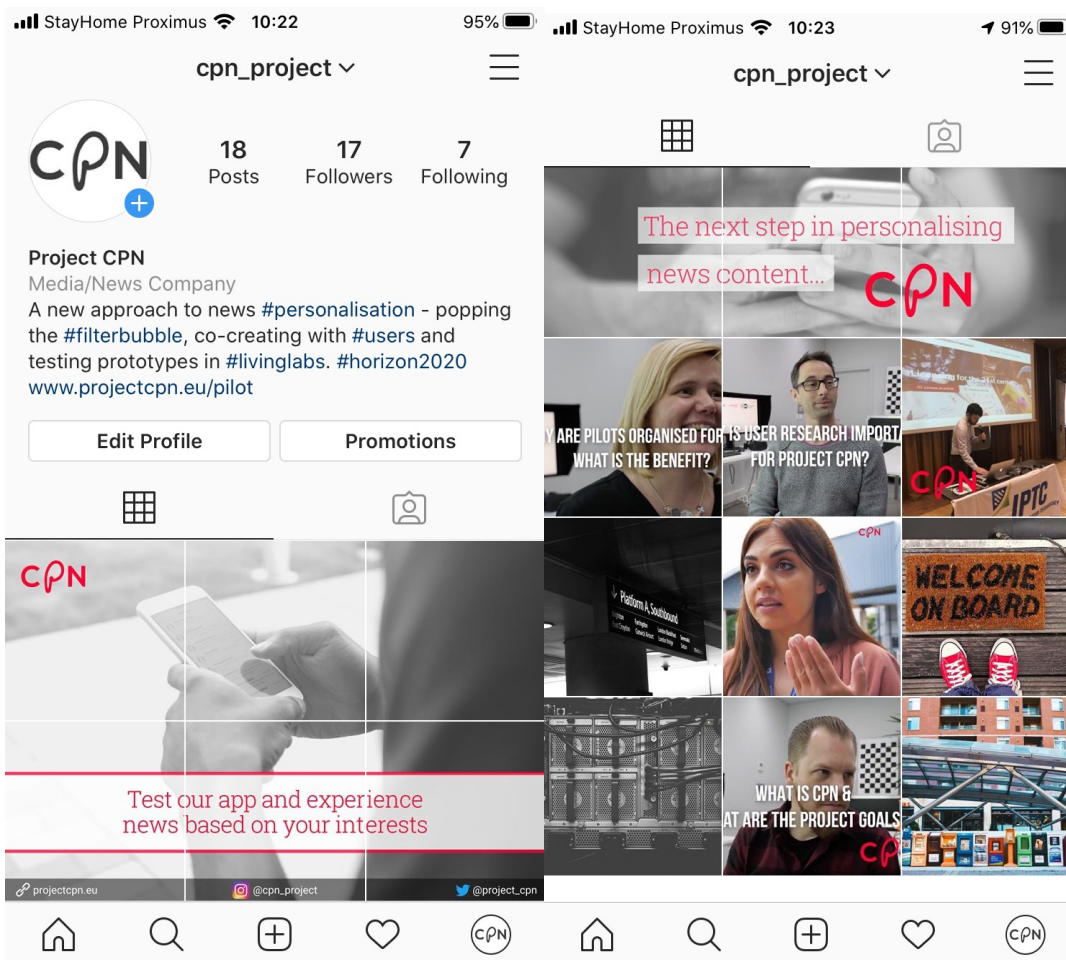
All project partners also used their own Twitter and other social media accounts to communicate about the CPN project.



1.3.2. Instagram account

We created an Instagram account for CPN (@cpn_project) in January 2020 for promotion of the 3rd pilot. The account was used mainly to be able to actively promote the CON pilot and as a landing page to attract and redirect potential testers to the pilot landing page on the CPN-website during the third pilot testing phase (Figure 2). To make it more attractive, we posted information about the project goals and set-up on the page as well. These posts were used to describe the technology bricks of the CPN-platform, to explain the project's view on news personalisation and privacy, to promote the CPN-newsletter, and to explain the importance and benefits of user research and why pilots are organised. Altogether 18 such posts were made on the Instagram-page during the period of 17-29 January 2020. The low number of followers (18 as of 30.04.2020) does not reflect on the purpose of the account, as the advertisement and function of the channel worked beyond the mere followership.

Figure 2: Screenshots of the CPN Instagram landing page

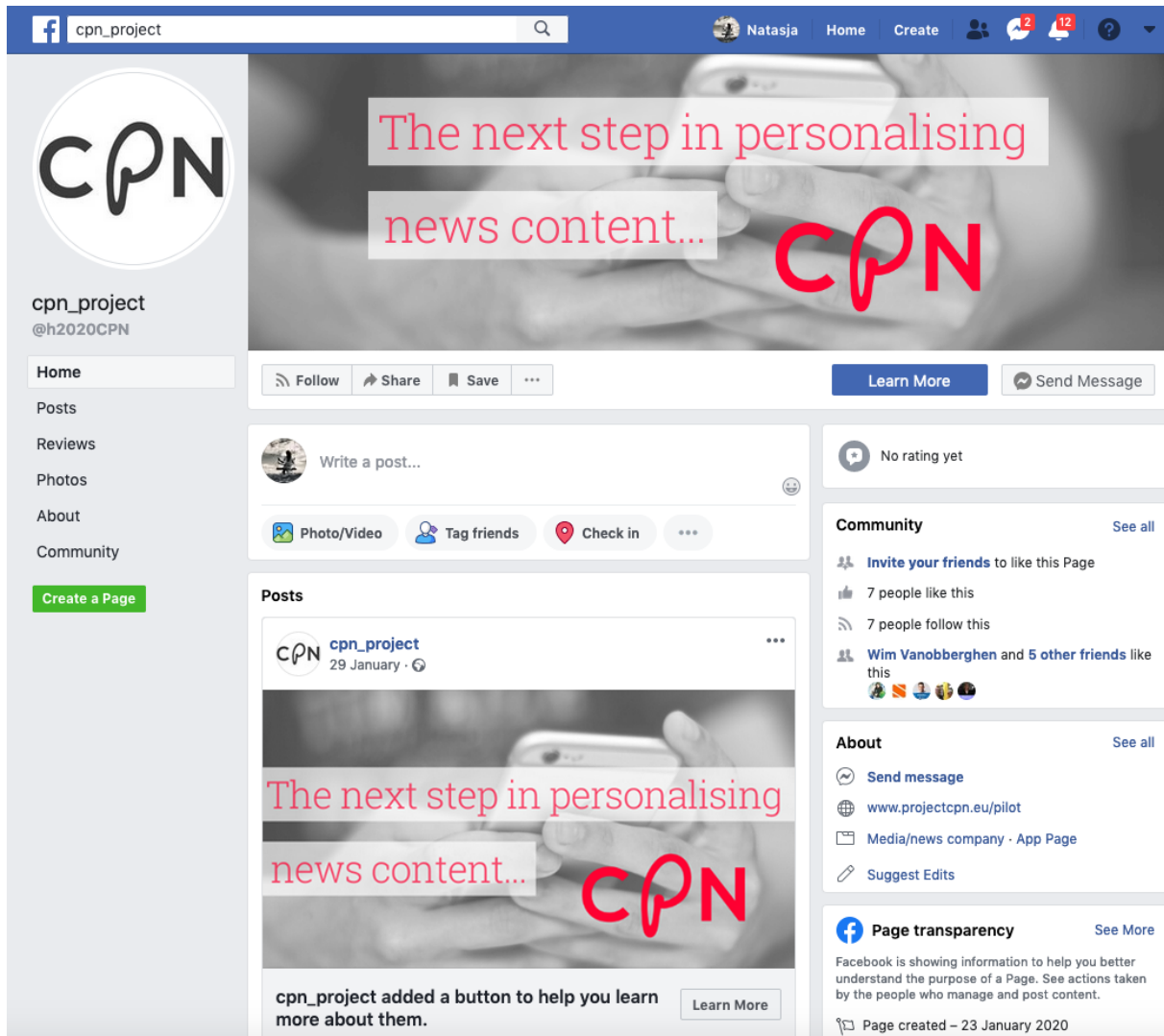


1.3.3. Facebook account

To further promote the 3rd pilot, we also set up a Facebook business page for CPN on 23 January 2020, cpn_project (@h2020CPN). Again the purpose of the page was mainly to be

able to set up a Facebook Ad account and link this account to the CPN Instagram account, with the aim of **running ads on Instagram and Facebook** - not to generate a large followership. The Facebook-page has a profile and cover photo, as well as an 'about'-description that redirects visitors to the pilot landing page on the CPN-website (Figure 3). We purposefully did not post any other content on the Facebook-page, which also reflects in the low number of likes the page received (7).

Figure 3: Screenshot of the CPN Facebook page

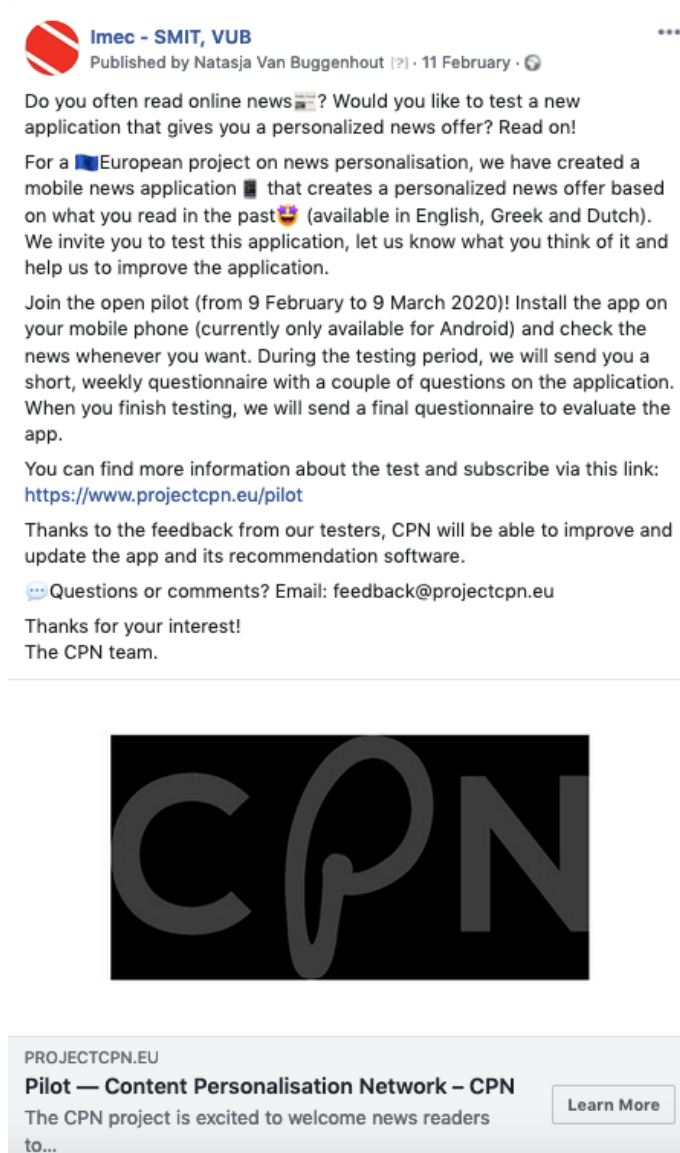


1.3.4. Social media campaign for pilot 3

We set up a social media campaign for the recruitment of end users during the third pilot testing phase. The campaign budget (net amount spent) was €300,46. Firstly, the social media campaign consisted of **boosted Facebook posts** (Figure 4). We boosted four Facebook-posts in total, one boosted post during each week of the CPN pilot 3 testing period. The overall objective was to promote the open pilot to a larger audience and stir people to the pilot landing page on the CPN website. We used detailed targeting to define our audience settings. Correspondingly, the Facebook-posts were targeted at people of all genders, ages

18 to 65+, who are located in Belgium, Germany⁷ and Cyprus. We targeted the Facebook-posts to people who access Facebook through (mobile) Android devices. In addition, our target audience was composed of people who indicated in their Facebook-profile that they are interested in Android apps, Android software development, digital media, digital trends, innovation, journalism, local news, mobile application development, news broadcasting, news magazines, online newspapers, personalization, technology, and newspapers.

Figure 4: Screenshot of a boosted Facebook post



Imec - SMIT, VUB
Published by Natasja Van Buggenhout · 11 February · Public

Do you often read online news? Would you like to test a new application that gives you a personalized news offer? Read on!

For a European project on news personalisation, we have created a mobile news application that creates a personalized news offer based on what you read in the past (available in English, Greek and Dutch). We invite you to test this application, let us know what you think of it and help us to improve the application.

Join the open pilot (from 9 February to 9 March 2020)! Install the app on your mobile phone (currently only available for Android) and check the news whenever you want. During the testing period, we will send you a short, weekly questionnaire with a couple of questions on the application. When you finish testing, we will send a final questionnaire to evaluate the app.

You can find more information about the test and subscribe via this link:
<https://www.projectcpn.eu/pilot>

Thanks to the feedback from our testers, CPN will be able to improve and update the app and its recommendation software.

Questions or comments? Email: feedback@projectcpn.eu

Thanks for your interest!
The CPN team.

CPN

PROJECTCPN.EU
Pilot — Content Personalisation Network – CPN
The CPN project is excited to welcome news readers to... [Learn More](#)

The boost duration for each Facebook-post was seven days. We spent €140 in total, €35 per week to boost each Facebook-post. The boosted Facebook-posts resulted in a total of **630 unique link clicks** to the pilot landing page on the CPN-website. The total reach was **97.664 users**. The total number of **impressions was 154.862**, with an average cost-per-

⁷ We opted for Germany as we wanted to use the DW reputation, which is high in Germany to attract participants that fitted our envisioned user groups.



click (CPC) of €0,25. The table below shows an overview of the schedule, budget, reach, impressions, CPC, and unique link clicks per boosted Facebook-post.

Table 8: Overview of the results of the boosted Facebook post

Schedule	Budget	Reach	Impressions	CPC	Unique link clicks
4-11 February 2020	€35	8.042	26.480	€0.38	88
11-18 February 2020	€35	16.972	24.422	€0.29	116
18-25 February 2020	€35	34.057	48.145	€0.17	200
25 February - 3 March 2020	€35	38.593	55.815	€0.15	226

Secondly, the social media campaign was composed of **Instagram advertisements**. We created two types of Instagram ads. The first ad was a sponsored post that was published on the **Instagram feed and explore page** (Figure 5). This ad ran from 4 to 11 February 2020. The ad objective, audience settings and detailed targeting were similar to the Facebook boosted posts (mentioned above). We spent €18,00, reached **46.184 users**, and generated **114.079 impressions**. There were **16 unique link clicks** and the average cost-per-click was €1,06. We decided to discontinue the ad after a duration of seven days because the results in terms of unique link clicks were too low in comparison to the CPC.

Figure 5: Sponsored post on the Instagram feed and explore page



The second Instagram ad was the format of an **Instagram story** (Figure 6). We ran the ad for the first time from 11 February to 3 March 2020. The ad objective, audience settings and detailed targeting were similar to the Facebook boosted posts (mentioned above). We spent €85,00, reached **46.184 users**, and generated **114.079 impressions**. This resulted in **212 unique link clicks** with an average CPC of €0,39. Since the results of this type of ad were positive, we decided to run the ad again but this time focused on recruiting women. Consequently, we ran the Instagram stories ad for a second time from 26 February to 5 March 2020. We spent €57,46, reached **72.584 users**, and generated **97.016 impressions**. This resulted in **234 unique link clicks** with an average CPC of €0,33.

Figure 6: Instagram story advertisement



1.4 PUBLICATIONS

In this section, we address the publications that have been published throughout the project. In the DOA, KPI 6.2 specifically mentions at least 20 publications, including scientific publications, popular publications such as blog posts, demonstrations and presentations. Looking at the broader numbers this target was reached.

For presentations held at different events see chapter 1.5.1 and for blog posts and newsletters see chapter 1.2. Below, we address the scientific publications and additional publications published throughout the project.

1.4.1 Scientific publications

There is only one academic partner in the project, which led to a lower number of scientific publications and a higher number of other types of publications. Imec did cooperate with the other project partners in some joint publications (some are still in progress):

1. Journal paper on the CPN dataset creation for information extraction from news articles, as kindly supplied by project partner Deutsche Welle. The articles are

annotated for NLP tasks including (i) named entity recognition, (ii) coreference and entity clustering, (iii) relation extraction, and (iv) entity linking. It is — to the best of our knowledge — the first work to combine all these tasks in the same dataset. Note that the dataset will be publicly available for research. Beyond the dataset itself, we also provide models to solve aforementioned tasks building on top of state-of-the-art. Tentative title and author list (paper to be submitted May 2020):

K. Zaporojets, J. Deleu, T. Demeester, C. Develder, *"DW-Articles: a multi-task document-level entity driven dataset for holistic information extraction"*, to be submitted May 2020.(see annex 1)

2. Abstract accepted for NEM 2020, Van den Broeck, W., Van Buggenhout, N., Lemmelijn, I., Mattheijssens, J., Van Lier, J. *"Have I got news for you! The challenges of news personalisation from an end user perspective"*. This paper focuses on the final user evaluation, the two-page abstract was accepted, the full paper is in preparation. authors are from imec and VRT. NEM was postponed due to covid-19, the new date of NEM2020 is still to be confirmed. (see annex 2)
3. Presentation at IAMCR 2019 Madrid, in which CPN was a case study. Authors: van Zeeland, I., Ranaivoson, H., Hendrickx, J., Van der Bank, J. & Van den Broeck, W. *"Measuring and profiling for online personalization in media while protecting personal data?"* This presentation is currently being transformed into a full article in which CPN will also be a case study.
4. Case-study on the evaluation by end users and professional users of the perceived transparency of the personal data receipt (PDR). Reporting will be in the format of a journal article. imec will take the lead, publication will be finalised in June).
5. Journal article on the evaluation of the CPN producer's dashboard by professional journalists (joint-publication by Cyprus University and imec - planned in May).
6. The CPN project and the personal data receipt are mentioned as an example of a communication approach that may increase consumer's trust in regard to media user data collection and processing in the following journal article:

N. Van Buggenhout, W. Van den Broeck, P. Ballon, *"Exploring the value of media users' personal information (PI) disclosure to media companies in Flanders, Belgium"*, to appear in *Mediální studia* (forthcoming May 2020).

1.4.2. Partners' publications

The CPN project partners also provided various other publications, published in different outlets. A first set of partner publications was already reported in D5.3. (10 publications). Since the last update, the following 7 additional publications were added to this list:

1. September 2019 : Trends in Newsrooms report / WAN-IFRA: An article about CPN was part of the "AI in the Newsroom" report <https://www.wan-ifra.org/reports/2019/09/30/trends-in-newsrooms-2-ai-in-the-newsroom>
2. September 2019: WAN-IFRA Magazine: Technology Guide & Directory: An article about CPN was included in the annual publication that features the latest technology trends, advice from experts, and a comprehensive directory of the industry's solution providers. <https://www.wan-ifra.org/articles/2019/10/28/tech-guide-directory-2019>
3. January 2020: "Our new technology exclusively for you! Read and win!" / DIAS: An article about the personalisation and the beginning of Pilot 3. <https://www.sigmalive.com/news/local/607900/i-nea-apokleistiki-texnologia-tou-sigmalive>



4. February 2020: "Everything you need to know on how to use our new technology" / DIAS: An article explaining to the user the process of the service My News on Sigmalive website
<https://www.sigmalive.com/news/local/611378/ola-osa-xreizetai-na-gnorizete-gia-tin-xrisi-tis-neas-mas-texnologias>
5. February 2020: "Sigmalive is innovating with the pilot version of its new technology" / DIAS: An article related to the interview of the General Manager, Christos Danezis, at Sigma tv's morning news broadcast "Protoselido" in which he presented the CPN pilot.
<https://www.sigmalive.com/news/local/613222/to-sigmalive-kainotomei-me-ti-dokimastiki-ekdosi-tis-neas-tou-texnologias>
6. February 2020: "VRT NWS zoekt testers voor gepersonaliseerd nieuwsaanbod" / VRT: An article to recruit testers for pilot 3 on the VRT website & the VRT NWS website.
<https://www.vrt.be/vrtnws/nl/2020/02/07/vrt-nws-zoekt-testers-voor-gepersonaliseerd-nieuwsaanbod/>
<https://www.vrt.be/nl/over-de-vrt/nieuws/2020/02/11/vrt-nws-zoekt-testers-voor-gepersonaliseerd-nieuwsaanbod/>
7. March 2020: "Innovative algorithms for transparent content personalisation" was written for CORDIS⁸ magazine (author: Joris Mattheijssens). It will be published soon, CORDIS will provide us with the full reference as soon as it is published.

1.4.3. Media mentions

Over the past 8 months, the project has been mentioned in the following external publications:

1. In July 2019, CPN was interviewed by Sophia Ignatidou from the "The Royal Institute of International Affairs" at the Chatham House in London. They are currently looking into AI and personalisation by legacy and social media. The paper was published in December 2019: AI-driven Personalization in Digital Media: Political and Societal Implications: CPN was mentioned in this paper about the implications of the adoption of AI by the legacy media as well as by the new media, focusing on personalization.
<https://www.chathamhouse.org/publication/ai-driven-personalization-digital-media-political-and-societal-implications>
2. In February 2020, the General Manager of Sigmalive, Christos Danezis (from DIAS), was hosted at Sigma tv's morning news broadcast "Protoselido" in which he presented CPN technology. <https://www.youtube.com/watch?v=Z9bCaut2yJo>

⁸ The Community Research and Development Information Service (CORDIS) is the European Commission's primary source of results from the projects funded by the EU's framework programmes for research and innovation (FP1 to Horizon 2020).



1.5 EVENTS

1.5.1 International conferences

During the CPN project, consortium members attended several international conferences, participating actively, through one or several of the following actions:

- Presentation of the project on the main conference programme,
- Exhibition booth or tabletop,
- Discussions/networking with relevant audiences.

A list of all events we attended during the project is listed in the tables below. CPN was present in 25 events. Unfortunately some events were cancelled or postponed due to travel restrictions related to covid-19. We also added a list of these events in which CPN would be represented otherwise.

Table 9 Events organised and attended during the final year of the project

Event	Date	Organizer	CPN activity	Attending Partners
84th Thessaloniki International Fair (TIF)	7–15 September 2019	HELEXPO	Booth presentation +	ATC
IPTC Autumn Meeting	14–16 October 2019	IPTC	Presentation	Dcat
NxtMedia	14 November 2019	NxtMedia WAN-IFRA	Presentation and panel	Dcat
Media Fast Forward	12 December 2019	VRT	Booth	VRT
Final CPN Webinar	7 May 2020	VRT WAN-IFRA	Webinar	ALL partners



Table 10 Events organised and attended during the second year

Event	Date	Organizer	CPN activity	Attending Partners
World Publishing Expo 2018	11 October 2018	WAN-IFRA	Stand	WAN-IFRA, ATC, IMEC
Newsroom Summit 2018	29–30 October 2018	WAN-IFRA	Presentation + table top	DW, ATC
Digital Media Asia 2018	8–9 November 2018	WAN-IFRA	Presentation	DW
ICT 2018	4–6 December 2018	European Commission	Booth	WAN-IFRA, VRT, DIAS, ENG
Cloud and Cyber Security expo	13 March 2019	CloserStill	Presentation	DCat
Cyber UK	24–25 March 2019	UK Government	Presentation	DCat
DW Global Media Forum	27–28 May 2019	DW	Booth + presentation	DW
World News Media Congress 2019	1–3 June 2019	WAN-IFRA	Presentation	DW, VRT, WAN-IFRA
IAMCR Conference, Madrid, Spain	7–11 July 2019	IAMCR	Presentation	Imec
4th JPEG workshop on Media Blockchain	16 July 2019	JPEG	Presentation	Dcat



Table 11 Events organised and attended during the first year

Event	Date	Organizer	CPN activity	Attending Partners
Investor Meeting for Media Innovator	5 October 2017	i3 H2020 CSA project	presentation	ATC
EC media cluster concentration meeting	16-17 October 2017	European Commission	presentation	ATC
News Impact	2-3 November 2017	European Journalism Centre & Google News Lab	Attendance + Networking	Digital Catapult
Multi-stakeholder conference on Fake News	13-14 November 2017	European Federation of Journalists (EFJ)	presentation	ATC
EBDVF2017	21-23 November 2017	Big Data Value Association	presentation	VRT, DIAS, ATC
2nd Media Lab Days	1-3 February 2018	Wan-ifra	presentation	VRT
DNIengage - use of machine learning or personalisation and recommendations in publishing	15 May 2018	Styria Media Group	presentation	VRT
Are we data? Identity, Journalism, democracy in the age of big data	22 May 2018	Technological University Cyprus	presentation	DIAS
DW Global Media Forum 2018	June 11-13, 2018	Deutsche Welle	Booth	DW
World Publishing Expo	11 October 2018	WAN-IFRA	presentation & booth table top	ATC, IMEC, WAN-IFRA



Table 12 Events planned during the final year of project, but cancelled or postponed due to Covid-19 travel restrictions

Event	Date	Organizer	CPN activity	Attending Partners
SxSW 2020	11 March 2020	SxSW	Networking	DCat
Digital Media Europe 2020, Vienna	1 April 2020	WAN-IFRA	Final event	WAN-IFRA, DCat, DW, VRT, SMEs (Yoop, U-Hopper, Loomi, Kensai)
NEM 2020	18-19th of May 2020	NEM	Paper presentation	imec, VRT

1.5.2 Workshops

During the first two years of the project, different workshops were organised (see D5.2. and D5.3).

In the third year (last 6 months), no additional user workshops were organised, because the focus was on the pilot 3 evaluation.

During Pilot 3 DW conducted 5 workshops with Journalism Master students and faculty members in 5 universities in Bangladesh. The goal of the workshops was to introduce the project and recruit participants for the pilot and first and foremost to introduce the topic of news personalisation and kick start a critical discussion about new technologies and their influence on traditional media with the students. This workshop series was conducted in cooperation with the DW Academy.

1.5.3 Hackathons

As mentioned in the previous deliverables, two successful hackathons were organised during year 2 of the programme (see D5.3). This led to the successful involvement of four



start-ups (Kensai, Loomi, U-Hopper and Yoop), who are still very active and involved with the consortium today. A full overview of their activities within CPN can be found in D5.5

1.5.4 Final CPN event (Webinar)

In order to end the CPN project and properly launch the CPN system our original plan was to organise a three hour event during WAN-IFRA's Digital Media Europe (DME) conference in Vienna on 1 April 2020. DME: <https://events.wan-ifra.org/events/digital-media-europe-2020>

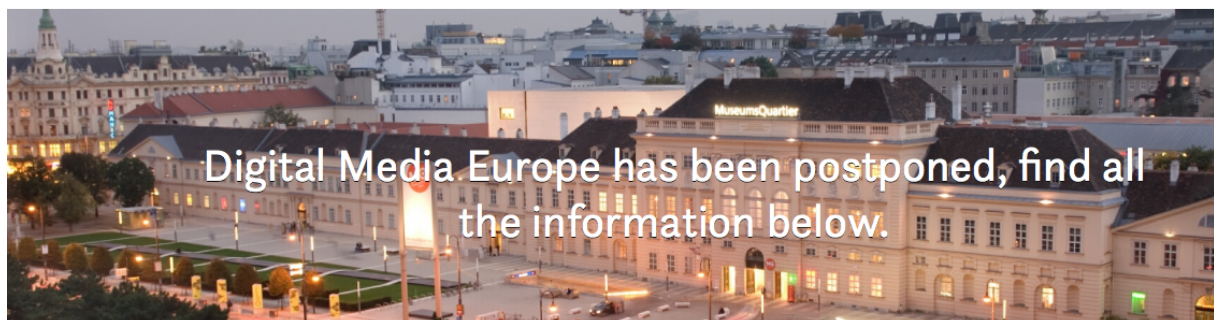
- A three-hour side session focused on personalisation (a final demo of the CPN tool and a selection of speakers on the broader topic of personalisation in the news industry, best cases, challenges, discussions on Privacy and transparency).
- an exhibition stand where participants can try the CPN app
- Exhibitions and presentations from the four SMEs (Kensai, Loomi, U-Hopper, Yoop) and their work for CPN

Unfortunately, due to the sanitary crisis in Europe, Digital Media Europe was cancelled and postponed to November 2020.

Figure 7 Digital Media Europe 2020 postponed

DIGITAL MEDIA | Europe 2020

10-11 November 2020, Vienna, Austria



As a consequence the Consortium decided in agreement with the Project Officer to organise a remote event in the form of a 90 minutes webinar around the topic of personalisation of news content after the end of the project: 7 May 2020.

On 30 April we had 60+ participants registered.

Subject – Webinar: Trust, Transparency and Personalisation – Building deeper relationships with your readers

The event will cover the following topics together with the panelists:



- What types of personalisation? What does personalisation mean for you?
- Why personalisation? What do we want to achieve?
- What are the effects in the current COVID-19 crisis?
- What about privacy and transparency?
- What are the challenges and how to overcome them?

Panelists:

- **Swantje Fischenbeck**, Innovation Management, Der Spiegel, Germany
- **Jarno M. Koponen**, Head of AI & Personalization, Yle News Lab, Finland
- **Gordon Edall**, Managing Director, Globe Labs, The Globe and Mail, Canada
- **Ine van Zeeland**, PHD Researcher on Privacy at imec-SMIT, Vrije Universiteit Brussel, Belgium

Spotlight on CPN startup partners:

- **Al Ramich**, Founder and CEO at Loomi.AI, UK: *'How to improve metadata for better personalisation?'*
- **Mattia Fosci**, CEO, ID-Ward, UK: *'How can media companies manage user identity and personal data in a privacy-first, post-cookie world?'*
- **Nicolò Pomini**, Software Engineer, U-Hopper, Italy: *'Solving the cold start problem?'*

The program will conclude with a Q&A session and a discussion with the audience. As a follow up, separate calls for demos of the CPN app and the SME solutions will be arranged on a per request basis. Of course the consortium will be open to demo CPN to any media company interested in learning more about it.



Figure 8 Webinar promotion on Twitter



1.6 OUTREACH TO EXTERNAL MEDIA COMPANIES

The description of work fixed an ambitious KPI concerning the engagement of external media organisations during the piloting phases: « Quantitative KPI 4.1: External organisations will be extensively involved in the piloting activities. It is expected that at least 8 external media organisations will test the platform during the project lifetime. CPN has already collected 5 letters of support. »

The main objective of these outreach and dissemination activities was to identify and secure potential early adopters and testers for CPN who pilot CPN technologies within their organisations. WAN-IFRA, imec and ENG have led the charge in organising and arranging these activities to keep momentum from the previous year going and to try and obtain early adoption and ensure experimentation with CPN microservices before the end of the product life cycle. Since mid 2019, the consortium has actively been engaging with external media companies in order to identify new partners to test the CPN solutions, if possible to pilot the tool using their own content and bring valuable feedback from a professional's perspective.

This is at the frontier between dissemination and exploitation. For confidentiality reasons, the details of the Publishers we have engaged and worked with are described in detail in D5.5 "CPN Business Models and Plans".

This section aims to explain our most important outreach activities to recruit new piloting media partners.

1.6.1 Outreach activities

Call for Interest (CFI)

In May 2019 we created a specific page on the website inviting external media companies to get in touch if they wanted to pilot the CPN recommender system.



Figure 9: Call for interest CPN website

LET'S DO IT.
BUT HOW?

Become a CPN pilot partner & personalise your content for **free!**

How you can join:

- 1. Explore CPN as a user**
Demo the platform as a visitor
- 2. Use the CPN platform with your own content**
Personalise your own news content in the existing CPN platform
Share the platform with your readers
- 3. Integrate the CPN software in your own app**
Personalise your news content in your own app with support from the CPN technical partners

CONTACT US

We promoted the page through the CPN newsletter on 7 May 2019, which proved to be quite successful. Following the mailing, the large regional French news publisher Sud Ouest got in touch to request a demo.

Followed in July an mailing to 3.300 media professionals (Product managers, Chief Digital Officers, Innovation Managers, Head of Digital, Chief Technology Officer, Research managers etc.) from the WAN-IFRA database, calling for testers. Ensued, around 15 expressions of interest and requests for more information.

We also disseminated the CFI regularly through the project’s Twitter account (sample tweet below), which was picked up and shared by project partners.

Figure 10 Call for interest CPN Twitter

projectCPN @project_cpn · Apr 25

Are you a news organization and looking to:

- ✓ #Personalise your news content
- ✓ Increase and please your readers
- ✓ Get GDPR compliant analytics and reliable sourcing of content

➔ Become a pilot partner of CPN! Find out more on projectcpn.eu

CPN Information Webinar 8 November 2019

On 8 November we hosted a CPN information webinar, and invited all the media companies that had expressed interest to know more to participate. The aim was, to give all the potential testers more information on the CPN project and explain the functioning of the tool and answer any questions they may have had.

The objectives of the call were to re-introduce CPN to media companies:

- goals of the project, features and functionalities of the tool, how companies could benefit from using it...
- Technical description: diving deeper, how CPN works, and how it can be implemented by media companies (what is needed from them to start testing)
- Inform the companies of what we expected from them (feedback collection)
- Inform them about the post-project plan: these media companies are potential future clients

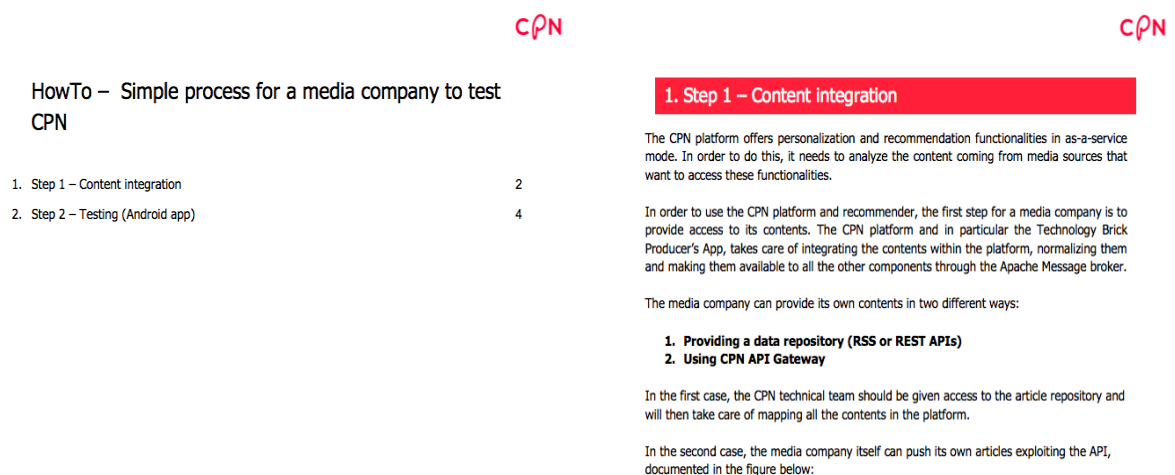
The webinar was attended by participants from Roularta (Belgium), GFR Media (Puerto Rico), 24 Sata (Croatia), AZ Medien (Switzerland) and Sunday Business Post (Ireland).

Distribution of a “How To Media Company Integration” document

As a follow up to the Webinar, all participants (including those who could not attend receive:

- a document explaining in technical terms How to integrate CPN for external media companies (see screenshot below)
- a link to a recording of the webinar
- the presentation slides

Figure 11: How to Media Company Integration document



The weeks and months that followed were mainly exchanges via email, further discussions with technical teams, answering any questions and understanding how to adapt CPN to each particular case.

1.6.2 Status of External media companies engagement (see D5.5)

To summarize, from mid 2019 the team officially started recruitment actions through mainly:

- open call for interest (website, emailings, personal invitations, social media etc)
- Engagement with external media companies (webinar and one-to-one demos)
- Nurturing relationships, understanding their needs and capabilities, matching them with the possibilities of CPN.

For more details of the work and relations developed with external media companies please see Section 2 of D5.5 "Exploitation supporting activities: outreach to external media companies"



CONCLUSIONS

This document provided an overview of our dissemination activities in the final phase of the project (September 2019 - April 2020) and reported on the CPN communication activities in relation to the communication KPIs set at the beginning of the project. Our communication and dissemination strategy consisted of a diverse set of cross-interacting activities, reaching out to varied audiences (media companies, SMEs, end-users and academia). The engagement of CPN's audiences via Twitter, offline discussions, representation at events, and a noteworthy list of publications, combined with other activities as outlined in this deliverable, show that the project reached a rather high level in terms of information diffusion and reputation building.

In addition, we have created promising links with external media organisations and had the opportunity to generate a substantial impact with the research and development results of the CPN project.

The current evaluation showed that the project achieved most of its goals regarding representation at and organisation of events and dissemination of the CPN project and results on the project website, social media and in various publications. The only areas where we did not completely reach our goals related to academic publications and website traffic. Some academic publications, based on the results of the final pilot and the dashboards analysis are still in progress, as the pilot results were only recently analysed. We are confident that they will be ready for submission soon.

Finally, the CPN dissemination activities followed the project's dissemination plan – with promising results – but inevitably the coronavirus situation in the Spring of 2020 caused significant complications for the final months of the project. We took action where possible to minimise the disruption: for instance, a final event was planned for April to present the CPN solution to European news editors and executives. After this was cancelled, we settled on organising a remote event.

Moreover, although 20 external media companies expressed interest in testing the CPN tool, the sanitary crisis understandably forced them to reorganise their activities and reprioritise their available time for research projects such as CPN. Because of this, many media companies could not ultimately commit to testing despite promising initial discussions. Many of them expressed interest in trying CPN at a later stage however, some providing a letter of interest.



ANNEXES**Annex 1: Draft publication imec ID-lab**

Noname manuscript No.
(will be inserted by the editor)

DW-Articles: A multi-task document-level entity-driven dataset for holistic information extraction**KLIM ZAPOROJETS · JOHANNES DELEU ·****Thomas Demeester · Chris Develder**

30 April 2020

Abstract In this paper we present DW-Articles, a multi-task dataset that combines four main Information Extraction (IE) annotation sub-tasks (i) Named Entity Recognition (NER), (ii) Coreference Resolution, (iii) Relation Extraction, and (iv) Entity Linking. Our main goal is to produce a dataset that describes interactions and properties of conceptual entities on a document level. This contrasts with currently dominant approaches that focus on describing specific entity mentions in text focusing on local interactions in sub-tasks such as relation extraction. By adopting document-level entity-centered approach, we are able to annotate entity properties that are not always explicitly stated in text and require additional reasoning step to be recognized. This presents an additional challenge when building models for DW-Articles since it requires an extra reasoning layer to derive the implicit meaning from the content of the document as a whole. Furthermore, unlike most of the IE datasets that are biased by either relying on pre-defined annotation schemas or on distantly supervised labeling, we follow a data-driven manual annotation approach. This produces a curated dataset with annotations that reflect the content of the corpus itself. Finally, we illustrate the advantage of using DW-Articles to train models in a joint learning setting. Concretely, we use graphical neural nets to jointly transfer the information from one task to another on DW-Articles,



outperforming this way independent single-task models.

Keywords Named Entity Recognition · Entity Linking · Relation Extraction

Corresponding author:

C. Develder
Ghent University – imec
iGent, Technologiepark Zwijnaarde 126
9052 Ghent, Belgium
Tel.: +32 9 331 49 61
E-mail: chris.develder@ugent.be



Information extraction (IE) tasks continue to receive ample interest, since they are key to unlock information embedded in a majority of data (especially on the web) that today still is unstructured. Such unstructured information can be leveraged maximally for further processing by converting it to structured data via IE. The majority of current IE works focus on a single specific IE task, such as: (i) named entity recognition (NER) [1, 3, 8, 30], (ii) coreference and entity clustering [9, 33], (iii) relation extraction [37, 48, 51], and (iv) entity linking [7]. Recently, such efforts that focus on a single task have been complemented with research that tries to take advantage of the inter-relation between tasks, thus trying to address two or more of them simultaneously and consistently. For example, [34] and [35] show state-of-the-art results when jointly learning named entity detection and coreference resolution (tasks (i) and (ii)). Similarly, combining tasks (ii) and (iii), [39] and [57] showed that a joint model trained on coreference and relation extraction tasks can achieve significantly better results than single-task approaches. Furthermore, [6, 54, 55] show an existing synergy between entity recognition and relation extraction tasks, achieving state-of-the-art results training on these two tasks (i) and (iii) jointly rather than separately. Also combining two tasks, [32] explores the dependency between entity linking and semantic relations between entities, and is able to train faster while achieving state-of-the-art results by including latent relation structural information in their model. Addressing the three tasks (i), (ii) and (iv) together, [28] demonstrates the superiority of a neural end-to-end architecture that is trained jointly on NER, coreference, and entity linking annotation layers.

However, despite this explosion of interest and success stories when using joint architectures, there has been limited effort in creating data-driven datasets that are jointly annotated for multiple tasks. Indeed, for each of the aforementioned IE tasks (i)–(iv), typically separate datasets are used to train and evaluate systems. Consequently, current models addressing multiple IE tasks together often use multi-tasking (with different datasets per task) rather than really joint modeling approaches.

A main objective that we address in the current paper is to fill that gap of an over-arching dataset covering all aforementioned tasks. We create DW-Articles, a new news-based dataset with annotations spanning 4 different semantic layers: (i) NER, (ii) Coreference Resolution, (iii) Relation Extraction, and (iv) Entity Linking. With a newly introduced annotation approach, we address four main limitations we observe in the currently prevalent datasets that are mainly exploited in a multi-tasking setup, as explained next.

First, we combine the stated 4 different tasks together. By addressing all tasks on the same corpus, DW-Articles contrasts with the existing well-known TAC-KBP datasets [14, 15, 23, 24, 25] where a different, non-overlapping corpus is used for each of the tracks.

Second, we use a bottom-up, data-driven approach where we manually define our annotations (in terms of, e.g., the entity and relation types) to reflect



the dataset at hand. Our motivation is that we want to maximally extract information of the corpus at hand (which we thus implicitly consider to be representative of data in the wild, given that we collect our news article corpus by random sampling; see further). Such bottom-up, data-driven annotation thus differs from currently dominant distant-supervision [20, 47, 61, 64] and top-down [14, 52, 53, 58] driven annotation approaches that are initially biased towards an a priori defined annotation schema (which is typically relatively limited, in terms of distinct entity and relation types).

Third, unlike currently prevalent sentence-based annotation practices [14, 20, 47, 58, 59], we rather use a document-level annotation approach for the entity clustering, linking and relation tasks. The motivation for such document-level annotation is to create robust annotations that are consistent across multiple sentences. Furthermore, this allows us to adopt a completely *entity-centric* approach where each entity (concept) is represented by one or more mentions located in a single or multiple sentences.

Finally, we not only annotate the elements (relations, clusters, entity types and Knowledge Base(KB) links) that are explicitly mentioned in text, but also the ones for which there is no direct evidence from explicit statements. Furthermore, we define a set of logical rules (see Appendix E) to (i) guarantee the consistency of different relations, and (ii) extend our set of annotations even further (e.g., if we know that a city *X* is located in a state *Y*, and that the state *Y* is located in country *Z*, we can also add a located-in relation between city *X* and country *Z*). This contrasts with the dominant purely evidence-based approaches [14, 20, 52, 58, 61] and introduces an additional common-sense reasoning evaluation component whose performance can be potentially boosted by using a world knowledge KB linking layer on top.

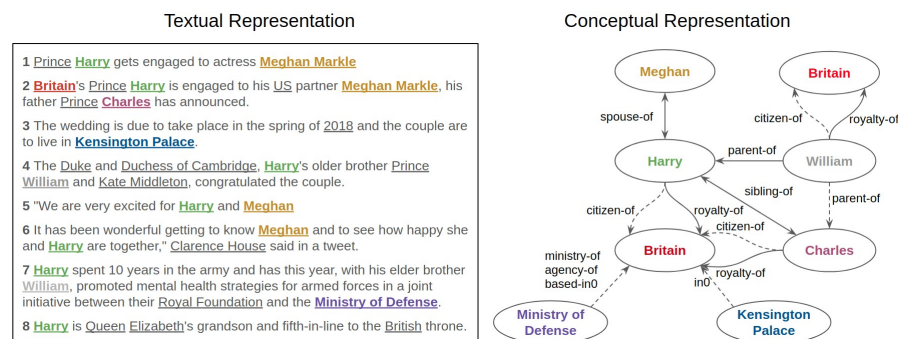


Fig. 1: An example from the DW-Articles dataset with entity mentions underlined. Out of a total of 29 annotated entities (concepts), we only show 8 (due to space constraints) to be represented in a graph on the right. This graph illustrates the relations that can be derived from the content of the article. The relations that are explicitly mentioned in the text (trigger-based) are depicted by solid arrows. Conversely, the relations that are implicit and/or need the whole document context (document-based) to be derived are represented by dashed arrows.



Figure 1 shows an example snippet from the DW-Articles corpus. As a first step, we identify all the named entity mentions in the text together with the possible entity types (e.g., “Harry” is a named entity mention of types *royalty*, *human interest*, *iptc::08000000*,¹ *person*, *keyword*, and *entity*). Second, we cluster all the mentions of entities in concepts on a document level. Thus, all mentions of *Harry* are clustered together, and similarly for the mentions *Meghan* and *Meghan Markle* since they both refer to the same concept. As a third step, we link the concepts to Wikidata, where we automatically apply additional clustering of entities that point to the same Wikidata entry. Fourth, we identify the keyword concepts in the article (e.g., Harry, Meghan). Finally, we identify the relations between the entity concepts. We mark both explicit (trigger-based) relations (e.g., spouse of *Meghan, Harry* ; parent of *William, Harry*) as well as implicit (document-based) relations (e.g., in *Kensington Palace, Britain* ; citizen of *William, Britain*) that are not explicitly stated in the text, but can be implied by the context or the world knowledge the annotator has. Such annotations were performed on 802 general news articles in English, selected randomly from a corpus collected from Deutsche Welle² between 2002 and 2018, as part of the CPN project³. The annotation process comprised three passes over the data, detailed further in Section 3: (i) annotation of entities and relations as identifiable by the annotator (i.e., not according to an a priori defined/limited set of entity and relation types), (ii) annotation based on the entity and relation types as identified in the first pass, (iii) parallel annotation by 2 annotators to assess inter-annotator disagreement, derivation of refined annotation guidelines and revisiting previous annotations.

Besides the dataset itself, we also contribute baseline models to address aforementioned IE tasks, inspired by state-of-the-art models developed in literature on related datasets. In particular, we experiment with newly proposed span-based approaches [34, 38, 39] and adapt the DyGIE architecture proposed in [39] in order to be used on the DW-Articles dataset. We jointly train NER, coreference and relation extraction layers and investigate the benefit of jointly solving those tasks versus stand-alone task-specific models. In evaluating the baseline model performance, we introduce a new evaluation measure to compute the relation extraction performance on DW-Articles. Current dominant systems consider *mention-level* scoring of relations when reporting on datasets such as ACE-2004 [5, 36, 62], ACE-2005 [39, 62], and TACRED [51, 63, 64]. In contrast, our proposed scoring algorithm is designed specifically to measure the relation extraction between entities that are composed by *multiple mentions* thus addressing two potential problems when using more traditional mention-based or cluster-based approaches: (i) a dominance of the score by relations between entities composed by many mentions as opposed to the relations between entities composed by single or few mentions that would present a purely mention-based approach, and (ii) the strictness of purely cluster-

¹ The IPTC types are universally defined news categories based on a media taxonomy (<https://iptc.org/standards/subject-codes/>)

² <https://www.dw.com>

³ <https://www.projectcpn.eu>



ter-based approach that, besides the relation type, also requires an exact match between connected entity clusters (i.e., mentions that compose predicted entities) in order for a predicted relation to be considered correct. We solve both aforementioned problems by designing a function to re-weight the number of predicted and gold relation links between entity mentions by the respective predicted and gold entity sizes. Our results suggest that, while challenging, DW-Articles opens new possibilities of research in the domain of information extraction by achieving superior results when using joint multi-tasking approach compared to single-task based models.

In summary, the main contributions of our work are:

- (1) We construct a self-contained dataset (Section 3) with joint annotations for four basic information extraction tasks (NER, entity linking, coreference resolution, and relation extraction), that provide document-level annotations (as opposed to typical sentence-level annotations for, e.g., NER).
- (2) We introduce a data-driven, bottom-up three-pass annotation approach complemented by context-based logical rules (Section 3).
- (3) We provide baseline models (Section 4) with source code for NER, coreference resolution, and relation extraction, and provide experimental results (Section 6) that demonstrate the potential of joint IE task models versus single-task solutions. For the relation extraction part, we additionally propose a new evaluation metric (Section 5), which is aligned with the entity-oriented nature of DW-Articles.

Information extraction (IE) tasks have been studied extensively over the last two decades. Traditionally, most of the IE datasets have focused on a single task, making it very challenging to develop systems that train jointly on different annotation sub-tasks of a single corpus. Thus, there are a number of very well known datasets for Named Entity Recognition such as CoNLL-2003 [49] and WNUT2017 [12], for Relation Extraction like Semeval-2010 T8 [21], TACRED [64], and FewRel [20], and for Entity Linking such as IITB [29], CoNLL-YAGO [22] and WikilinksNED [16]. Conversely, in this work we propose a multi-task dataset as a single corpus annotated with different information extraction layers: named entities, entity clustering, relations between clusters of entities, and entity linking as our main core annotated tasks. We further complement our dataset with additional tasks such as document classification and keyword extraction. Table 1 presents an in-detail *qualitative comparison* between DW-Articles and some of the most well known related datasets, while Table 2 provides *quantitative corpus statistics*.

Most prominent efforts to produce jointly annotated datasets have focused on using a *top-down* annotation approach. This method involves an a priori defined annotation schema that drives the process of selection and labeling of the corpus. Consequently, the resulting dataset consists of artificially selected documents that contain a minimum amount of previously defined annotation



Dataset	Core				Doc				Ent				Unaide				Open
	NER	Coreferen	Relatio	Linkin	Coreferen	Relatio	Multi-label R	Keyword	Classificati	Multi-label Ent	Relatio	Linkin	NER	Coreferen	Relatio	Linkin	
DW-Articles	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BC5CDR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TAC-KBP 2019	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SciERC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DocRED	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TAC-KBP 2015	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TAC-KBP 2017	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MUC-7 (1997)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
OntoNotes 5.0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ScienceIE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Rich ERE (2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TAC KBP 2010 (EDL)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GENIA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
FewRel	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ACE 2005	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NYT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ACEtoWiki	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AIDA CoNLL-YAGO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SemEval 2010 T8	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MUC-4 (1992)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
FrameNet (2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Qualitative comparison of the datasets. We divide our comparison in five groups: (i) *Core Tasks* represent the main sub-tasks covered in DW-Articles, (ii) *Doc Level* indicates different sub-tasks annotated on document level, (iii) *Ent Level* indicates which annotations are done with respect to entity clusters as opposed to individual mentions, (iv) *Unaided* specifies if the annotation process was completely manual or some form of distant supervision involved, and (v) *Open* indicates whether the dataset is freely available.

types. The de facto datasets used in most of the joint learning baselines such as ACE 2005 [58], TAC-KBPs [14, 15, 23, 24, 25] and Rich ERE [52] use this annotation approach. More specifically, during the creation of ACE 2005 dataset [42, 53, 58], the annotators initially tagged candidate documents as “good” or “bad” depending on the estimated number and types of entities present in each one. In subsequent annotation stages, only “good” documents were fully annotated and included in the final dataset. Similarly, during the creation of TAC-KBP [14, 15] datasets, the annotators focused on creating evenly distributed annotations between three entity types (PERs, ORGs, and GPEs), as well as required the candidate documents to contain a minimum number of entities related to event types. In case of Rich ERE [52], the documents to annotate were prioritized by the event trigger words density calculated per 1,000 tokens, focusing thus only on content with a high number of previously defined key event-related tokens. Furthermore, other IE-related datasets [2, 20, 61, 64] use similar pre-filtering techniques in order to select the text to be annotated. As a consequence, the corpus and annotations in these datasets are biased and not representative of the language used in the different input domains. Conversely, we adopt a radically different *bottom-up* approach where we derive the annotations (e.g., entity classification types, relation types) from the data



itself. This bottom-up data-driven procedure, guarantees that the resulting dataset is representative of the document corpus and reflects the particularities of the language used in its journalistic domain. Furthermore, it is better suited to reflect the properties that are inherently present in written corpora such as long-tail distribution of different annotation types.

Driven by data-hungry models, there has been a growing interest in making use of *distantly supervised* algorithms to either automatically produce big datasets from scratch [43, 45, 47], or to provide an assistance tool to make the annotation process easier and faster [20, 61, 64]. For example, with 5M tokens, the NYT dataset [47] is the largest corpus in the comparison Table 2. It was produced completely automatically by using a factor graph model with an *expressed-at-least-once* assumption, i.e., assuming that two related entities appear at least once in a single sentence. The authors used OpenNLP for NER and Freebase as external supervision KBs to derive relations. Conversely, when building FewRel [20] dataset, the crowdworkers were given sentences with distantly annotated entities (using spaCy) and relations (taken from Wikidata) and had to filter out the incorrect ones or the ones that couldn't be derived from the sentence semantics. Only the relations connecting more than 1,000 instances in the corpus were taken into account. Following a very similar procedure in the DocRED [61] creation, automatically extracted named entities were first linked to Wikidata using exact match heuristic and then connected with respective Wiki relations. Afterwards, an additional human annotation pass was performed to validate these relations. While capable of producing large datasets using very few annotation resources, these distant supervision methods rely on a set of initial assumptions that result in a lack of annotation quality. First, the use of automatic NER tools (e.g., OpenNLP or spaCy) can introduce errors. Second, the existing KB relation types (e.g., derived from Freebase or Wikidata) used for annotation, bias the final dataset towards a preliminary established domain. Third, the use of heuristics (e.g., exact match heuristic for entity linking in DocRED) can potentially introduce further annotation mistakes. Finally, these datasets are restricted by preliminary established constraints such as the total number of covered relation types or the intra-sentence assumption in the FewRel and NYT datasets. As a consequence, the resulting annotations are limited in coverage and only focus on a rather constrained number of previously defined cases. We address these shortcomings by deriving our own annotation schema manually, completely from scratch starting from the corpus itself. This way, we are able to discover a wide variety of annotation types that can maximally cover the meaning representation of our corpus. Furthermore, we ensure the quality of our annotations by following a three-pass procedure where in the last pass we perform an independent parallel annotation of our dataset and correct the mistakes by carefully analyzing each of the found discrepancies.

Finally, from the perspective of the necessary evidence to annotate a particular entity type or relation, the currently existing datasets can be divided in *trigger-based* [14, 15, 20, 21, 23, 24, 25, 38, 52, 58, 64] and *document-based* [18, 26, 45, 46, 55, 61] annotations. The trigger-based datasets require that



Dataset	Tokens	Mentions	Clusters	# NER types	# Unique Rel. instances	# Relation mentions	# Relation types	# Mention KB links	# Unique KB links
DW-Articles	501,095	43,383	23,133	331	21,580	226,102	69	28,484	13,078
ACE 2004	185,696	29,949	12,507	43	5,525	5,976	24	-	-
ACE 2005	259,889	54,824	37,622	51	7,786	8,419	18	-	-
MUC-4 (1992)	717,798	14,196	-	13	-	-	-	-	-
SciERC	65,334	8,094	1,015	6	-	2,687	7	-	-
OntoNotes v5	2,088,832	161,783	136,037	-	-	-	-	-	-
GENIA	554,346	56,743	?	5	-	2,337	2	-	-
CoNLL-2003	301,418	35,089	-	4	-	-	-	-	-
DocRED	1,018,297	132,392	98,610	6	50,503	155,535	96	-	-
CoNLL-YAGO	301,418	-	-	-	-	-	-	34,929	5,599
TAC-KBP 2010 (EDL)	3,053,336	6,495	3,750	-	-	-	-	3,818	2,094
FewRel	1,749,300	?	140,000	-	70,000	?	-	?	140,000
SemEval 2010 T8	207,307	21,434	-	-	-	6,674	9	-	-
NYT	5,765,332	1,388,982	-	-	-	142,823	52	1,388,982	69,506
TACRED	?	-	-	-	-	21,784	42	-	-
WNUT 2017	101,857	3,890	-	6	-	-	-	-	-
BC5CDR	343,175	29,271	10,326	2	3,116	47,813	1	29,562	2,351
ScienceIE	99,580	9,946	9,536	3	-	638	1	-	-

Table 2: Numerical comparison between the datasets. Note that the same entities, as well as relations between them, may be expressed more than once per document. Hence, we provide both the total number of entity mentions linked to a KB entity, as well as the unique number of such links; similarly for relation mentions and their unique count.

a particular relation or NER type should only be annotated if it is supported by an explicit reference in a text. For example, in Fig. 1 there is a concrete reference of the relation between “Meghan” and “Harry” in form of triggers such as “gets engaged” in sentence 1 and “The wedding” in sentence 2. Most of the traditionally used jointly annotated datasets such as ACE 2005 [58], TAC-KBPs [14, 15, 23, 24, 25] and Rich ERE [52] as well as others such as FewRel [20], OntoNotes [59], TACRED [64], SemEval 2010 Task 8 [21] and SciERC [38] are *trigger-based*. The disadvantage of such an approach is that it only captures the most simple cases of relations and NER types that are explicitly mentioned in the text. As a general rule, this also limits the datasets to cover only the relations between entities that appear within a single or at most few adjacent sentences where the relation trigger appears. However, as we move to a broader *document-based* interpretation, it is common to find relations that are not explicitly mentioned in text. Thus, in our example of Fig. 1 the relation between “Ministry of Defence” and “Britain” is not explicitly indicated in the text. However, after reading the whole article we can infer relations such as “ministry of”, “agency of” and “based in” between these two entities. Unfortunately, a manual document-based annotation process is a time-consuming task. As a consequence, most of the authors have resorted to the use of partially [61] or completely [26, 45, 55, 61] distantly supervised annotation pipelines to obtain such datasets. In contrast, we create the relation and entity type annotations manually. This allows us to produce a cleanly annotated dataset by following specific annotation guidelines. The only other dataset in using completely manual approach to extract document-based relations we are aware of is MUC [18, 46]. However, it is limited to only 3 entity and 3 relation types (MUC-7 version) compared to 331 entity and 69 relation types in our dataset. Moreover, we define additional 843 additional logical rules on entity tag level and 41 logical rules on relations (see Appendix E) to check some reasoning patterns in the data. These rules range from a simple implications between NER tags such as *type* tag “war” also



implies the *topic* tag “conflict”, to a more complex logical reasoning such as in example of Fig. 1 where from the text we know that “Harry” is a “sibling of” “Charles”, and “William” is a “parent of” “Harry”. However, a further logical rule $siblings(X, Y) \wedge parent(Z, Y) \Rightarrow parent(Z, X)$ allows us to automatically check for another relation of “parent of” between “William” and “Charles” that is implied in the article. This way, we are able to inject additional level of common-sense knowledge into DW-Articles.

In this work we introduce a *bottom-up* data-driven annotation approach. Our main goal was to have the annotation schema reflect most, if not all, of the types of entities and relations that are effectively mentioned throughout the corpus and hence be able to structurally extract most of the information embedded in the corpus. This is why we derived the annotation schema from the corpus itself, for which we organized the annotation process in three passes, that are described in more detail below: (i) *exploratory*, (ii) *schema-driven*, and (iii) *inter-annotator refinement*.

3.1 Exploratory pass

During this pass, our main goal was to discover the annotation structure (i.e., annotation schema) to be used on the corpus. Three annotators were involved in this step: one expert annotator and two job students. However, no parallel annotation was done and the role of the expert annotator was to annotate part of the corpus himself as well as instructing and supervising the job student annotators. No a priori fixed schema was followed, but we required annotators to be as consistent as possible during the process. Figure 2 illustrates this concept on relation-level annotations. In input text box in the bottom right, the annotators can freely define, modify or delete relations and relation types between entities by following a previously defined syntax. The same procedure is also used to annotate coreferences, entity linking, entity types, and keywords.

Below, we describe more in detail the chronological steps followed during this exploratory pass.

Document collection — We began with a complete collection of 156,607 articles of Deutsche Welle⁴ (DW) spanning a period from 2002 to 2018. Out of this set, we randomly selected 804 articles to be fully annotated.

Topic annotation — We proceeded to annotate each of the documents with topics derived from the IPTC⁵ news media standard. The annotators were asked to assign at least one IPTC topic to each of the news articles.

Keyword annotation — In this phase, the annotators were asked to annotate the main named entities that could best describe (were relevant) to a

⁴ <https://www.dw.com/>

⁵ <https://iptc.org/>



Taking stock of press freedom in Germany¹⁷

Fifty years ago, the offices of **Spiegel** magazine were searched and leading journalists arrested as a result of political pressure. Press freedom eventually prevailed. Today too the media's independence is under threat.

What happened in October, 1962⁰, had a lasting effect on the relationship between politicians, the judiciary and journalists in Germany¹⁷. **Spiegel** magazine had published an article critical of Germany¹⁷'s armed forces, the **Bundeswehr**⁴. The **German**¹⁷ army, it said, was only "prepared for defense to a limited extent", adding it would certainly have to employ nuclear missiles to counter attacks. That was the author's conclusion following an evaluation of a NATO⁰ maneuver. Leading politicians - first and foremost Chancellor⁰ **Konrad Adenauer**⁴ and Defense Minister⁰ **Franz Josef Strauss**⁴ - accused the magazine of treason. A constitutional lawyer filed charges against the editorial office, arrest warrants were issued for the author and the chief editor. The latter remained in custody for more than 100 days. The **Spiegel**² editorial department was shut down for weeks. There was a public outcry against the state's handling of the journalists, even protest from members of the coalition government. In the end, Defense Minister⁰ **Strauss**⁴ was forced to resign after his considerable participation in the reprisals against the news magazine became evident. Germany¹⁷'s **Federal Court of Justice**⁴ also ruled there was no proof of treason. What became known as the **Spiegel Affair**¹ still touches issues that are highly relevant even today. Is a democratic state allowed to reprimand critics reporters? And how freely can the **German**¹⁷ media report on politics? Press freedom today: Such featherbrained attempts by politics to influence the press are today no longer possible". **Bodo Hombach**³, former **Social Democrat**¹ and director of **WAZ**⁴ media group, told **Deutsche Welle**². The head of the **German Federation of Journalists**⁴ (**DJW**⁴), **Michael Konken**² agrees. "We have a critical media with quality publications and broadcasters who bring such problems to the public." At the same time, **Konken**² has to admit that **Reporters without Borders**⁰ has in the past years ranked Germany¹⁷ only in places 17 to 19 in its press freedom index. The top places were occupied by the Scandinavian⁰ countries. This, **Konken**² said, pointed to problems in Germany¹⁷. And indeed, there have been examples confirming that in the recent past. Politics has grip on public broadcasters. The **Breider case**¹ is often viewed as an example of a blow to freedom of the press. **Nikolas Breider**⁴ was editor-in-chief of Germany¹⁷'s public broadcaster⁰ **ZDF**⁴. As he was an independently minded journalist who would not refrain from saying what he thought, he was not always popular with politicians. It was first and foremost **Roland Koch**², premier of the state of **Hesse**⁴ at the time, who opposed renewing **Breider**⁴'s contract - and **Breider**⁴ lost his job. How was this possible? All public broadcasters are officially controlled by their supervisory and administrative boards. After the experience of government controlled media under the Nazis⁰, broadcasters in post-war **West Germany**⁴ were supposed to be democratically controlled. The boards are made up of a cross-section of society, including politicians from all major parties. In the case of **ZDF**⁴ when **Breider**⁴ was editor in chief, the board was dominated by conservative politicians. And that's what stacked the odds against **Breider**⁴ when it came to extending his contract. Powerless the way in which **Breider**⁴ was treated by the politicians on the board caused an uproar across the entire media landscape in Germany¹⁷. Even representatives of the conservative media thought that a line had been crossed. "An act of unbelievable high-handedness" of those responsible, says **Matthias Döppner**², who is at the helm of the conservative **Axel Springer**⁴ publishing house. "A dangerous interference with the DNA of a broadcaster," criticized **Frank Schirrmacher**¹ of the **Frankfurter Allgemeine Zeitung**³ daily. But, again and again, there have been attempts - mostly by politicians - to interfere with the editorial affairs of a broadcaster's claim confirmed by **Hartmann von der Tann**⁴, former coordinator for politics with Germany¹⁷'s other big public broadcaster⁰, **ARD**³. Print media also affected Newspapers in Germany¹⁷ experience limits to press freedom for different reasons. **Dirk C. Dieck**¹, for his book⁰ **The Fourth Power**¹, spoke to 25 leading journalists and editors-in-chief about how independent they are. There's not a lot of room, he says. Speaking to **BWZ**², the author describes the sobering answers he got and how almost all of the journalists describe the same vicious circle: it began with the boom of the Internet⁰. More and more people got their news from the Web⁰; in particular, young people who are important for advertising. Many companies therefore shifted their advertising budgets to focus on the Internet⁰. Germany¹⁷ publishers responded by setting up online newsrooms to get some of the money going into online ads. But the free content offered online has not generated enough revenue through advertising. The print media has failed to capitalize on the online market. With the same budget, publishers have since tried to sustain both their print and online editions. As a result, editorial offices have been split, and more and more work ends up on fewer shoulders. The increase in the workload and the Internet's hunger for new stories within hours, or even minutes, puts enormous pressure on journalists. Proper independent research has become a luxury. Risko manipulation "What dominates is content that can quickly be generated," says **Dieck**¹. It is rare that investigative research really uncovers a political scandal. Instead, there are quick party stories or the quick report on a police patrol catching *cracklin thrushes*. he notes: "Content is neither challenger nor more tabloid-like, while a lack of time and money means journalists follow the herd"

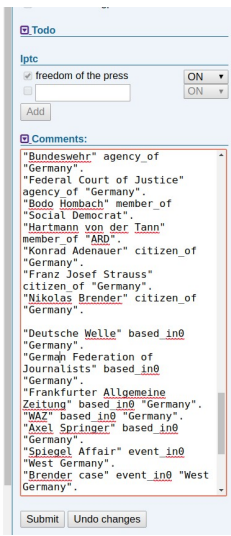


Fig. 2: Visual interface used for annotation. The annotators are given freedom to define their own annotation types such as relations in the input text box in the right corner.

particular news article. We set no hard limit to the number of main entities to annotate. However, we advise it to not be higher than 5.

Entity annotation — All the named entities are manually annotated by the expert annotator. A set of tags is assigned to each of the entities describing its properties such as type, topic, IPTC code, gender (if applicable), etc. The focus is to identify a set of categories that are repeated across the documents.

Entity coreference — The entity coreference was done on named entity level without involving nominal and anaphoric expressions. The entity mentions referring to the same concept are grouped in clusters. This makes the dataset highly entity-oriented since all the annotations are defined in relation to entities, each of which cluster one or more named mentions. The process of entity coreference was carried out by the expert annotator.

Entity linking — Once all the entity mentions have been identified and coreferenced, we proceed with the entity linking process to the Wikidata⁶ knowledge base (KB). For entities that do not appear in English version of the KB, we first try to find a match in other wiki languages and, if not found, consider the link as NIL. The annotations are carried out on mention level and are also used to complement the coreference clusters created in the previous step. This is done by automatically grouping mentions of the same entity (i.e., those mentions linked to the same Wikidata entry) in a single concept cluster. Table 12 in Appendix C shows the numerical statistics on linked entities for our final version dataset.

⁶ <https://www.wikidata.org/>



Tag Category	# Ents	% Ents	# Mens	% Mens	# Classes	Labels per Entity
type	21761	94.1%	43139	99.4%	190	2.9
topic	7850	33.9%	18363	42.3%	14	1.0
iptc	7066	30.5%	17199	39.6%	113	1.3
slot	3001	13.0%	14323	33.0%	7	1.0
gender	407	1.8%	849	2.0%	2	1.0
TOTAL	23133	100.0%	43383	100.0%	333	3.8

Table 3: Main entity types where *Ents* and *Mens* stands for *Entities* and *Mentions* respectively.

Relation extraction — With all our entities identified, linked and coreferenced, we proceeded to annotate the relations between them. In this first exploratory pass we instructed the annotators to be creative and think in different meaningful ways two entities from text could be related. Unlike currently dominant datasets where the relations are annotated on the level of each of the individual entity mentions (mention-based relations) [2, 14, 15, 23, 25, 27, 38, 52, 58, 64] triggered by specific expressions, we focus on annotating relations between entities themselves (concept-level). This allows us to think concept-wise and come up with relations that are not only *explicitly*, but can also be *implicitly* derived from the text. As a result, our dataset includes predominantly long-range relations, i.e., between entities that involve mentions located far away (i.e., multiple sentences away) from each other on a document level. Figure 3 compares the *minimum* distance between two mentions involved in relation for various mention-based relations (Rich ERE⁷, TAC-KBP⁸, and ACE 2005) and entity-based relation (DocRED, BC5CDR, and the final version of our DW-Articles dataset) datasets. The x -axis represents the distance in tokens (left) and sentences (right) and y -axis denotes the percentage of covered relations in a dataset. We can immediately observe that the relations in DW-Articles dataset involve mentions that are located further apart. Furthermore, we also observe how other datasets that define the relation in terms of entities (BC5CDR and DocRED) require larger token and sentence spans to cover all the relations in the respective dataset. This is explained by the fact that entity-based relations very often involve mentions located in different sentences in the document that refer to those entities. This is not the case for mention-based relations such as in TAC-KBP, Rich ERE and ACE 2005 datasets, where the annotation bias is towards finding explicitly mentioned relations in text between concrete mentions of entities, often within a single sentence.



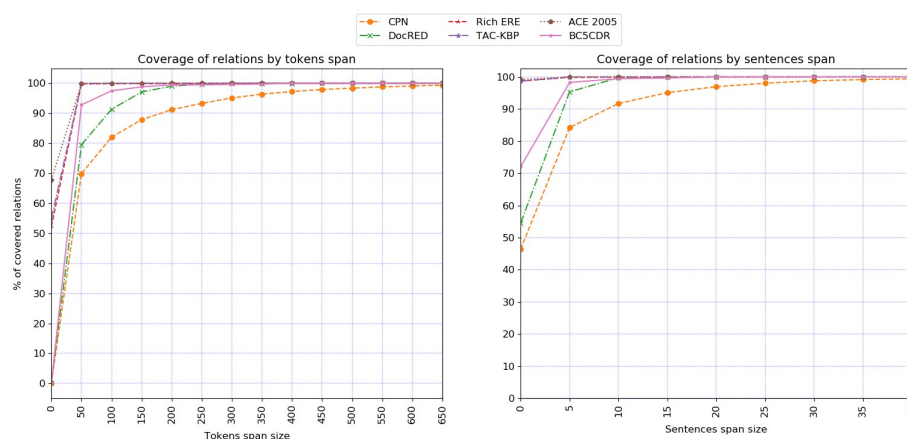


Fig. 3: Comparison of the coverage of the number of relations by token span size (left) and sentence span size (right). It can be observed that the average distance between the related mentions in our dataset is higher than of other widely used datasets.

3.2 Schema-driven pass

The main goal of this step is to create a consistent annotation schema for (i) named entity and (ii) relation types based on the annotations made in the *exploratory pass*. As a first step, we identify the multi-label tags to be assigned to **entities**. (The next step for relations will be discussed subsequently.) We divide entities in five main categories:

1. *Type*: The derived entity types fall in one of three main categories: *Entity* (i.e., named entity proper nouns such as "Germany"), *Values* (*money, time, and role*), and *Other* (e.g., adjectives referring to entities such as "German"). Each of these entity types can hierarchically be expanded to more granular sub-types. For example, one of the sub-types of *Entity* type is *person* (see Table 10 in Appendix A), which itself recursively expands into 37 sub-types: *religion clergy, education teacher, politician, etc.* The advantage of defining this named entity types taxonomy is that we can easily extend it and add additional sub-types. For example, if we observe a big number of annotated entities under the *entity person politician minister* sub-type, we can quickly filter the respective mentions and visually determine whether it would be reasonable to add additional sub-types (e.g., *minister of finance, minister of education, etc.*). This method gives a substantial speed boost in our annotation process and goes hand in hand with data-driven bottom-up annotation approach where we add entity sub-types only if we observe they are actually reflected in the data.
2. *Topic*: We categorize our named entities in different topics. Table 11 in Appendix B provides a detailed illustration on how the topics interact

⁷ We use the Rich ERE dataset from the LDC2015E29 and LDC2015E68 catalogs.

⁸ We use the TAK-KBP 2017 dataset from the LDC2017E54 and LDC2017E55 catalogs.



# Relation Labels	Related Ent. Pairs	%Related Ent. Pairs	# Related Men. Pairs	% Related Men. Pairs
1	12679	76.05%	111537	69.18%
2	3105	18.63%	34933	21.67%
3	884	5.30%	14650	9.09%
4	3	0.02%	100	0.06%
TOTAL	16671	100.0%	161220	100.0%

Table 4: Relation multilabel stats

with our entity types. This tagging approach enriches a more traditional entity type annotation based on topic agnostic types present in datasets such as CoNLL-2003, ACE 2005, TAC-KBP and Rich ERE. The advantage of linking topics to entity types is that it can be used as an extra feature to perform document-level topic modeling. Additionally, the topic information can also be used to disambiguate and further boost the performance on entity linking [17, 19].

3. *IPTC* : we use IPTC’s main generic categories (e.g., politics, sport, religion, conflict, etc.) to classify our entities. Furthermore, we map our entities to more specific IPTC sub-categories where we found a good categorization match with our dataset, namely for *sport*, *sport event*, *business sector* and *policy*.
4. *Slot* : represents additional categorization that is transversal to different entity types. One example is *interviewee* that can be assigned to any person interviewed in a particular article. Other possible slot values are: *keyword*, *in*, *head*, *death*, *interviewer*, and *expert*.
5. *Gender* : the gender in case of entities that refer to people.

The coverage of entities as well as the number of classes in each of these entity tag classes is represented in Table 3. We observe a combination of different characteristics that makes our DW-Articles entity classification potentially more challenging than in other similar datasets. First, the number of labels is rather big compared to other similar datasets (check Table 1 for details). Furthermore, our *type* tag is organized in a hierarchical structure making it more challenging to predict more granular sub-types. Second, we recognize that the entity classification is multi-label by nature and allows different complementary types to be assigned to a particular entity. As listed in Table 3, the average number of labels per entity is 3.8. This contrasts with prevailing single-label multi-class datasets such as ACE 2005[58], TAC-KBPs [14, 15, 24, 25], Rich ERE [52], WNUT 2017[12] and CoNLL-2003[49]. Third, from Table 10 (see Appendix A) we observe that the tag distribution is very unbalanced, where some types appear more frequently than others in our dataset. This opens the possibility to explore zero-shot and few-shot learning techniques using DW-Articles dataset.



The annotation schema consolidation process for **relation annotations** is very similar to named entities. We aim to create relation types that also depend on the entity *tag* types of the connected named entities. The reason for this tag dependency is that we expect that the joint models can benefit from the resulting cross-task type inter-connection. Thus, in the example of Fig. 1 we have a relation between a named entity sub-type *ministry* and a location sub-type country represented as *gpe0* in our annotation schema. The hierarchical type path to *ministry* is *entity* \rightarrow *organization* \rightarrow *gov org* \rightarrow *ministry*. In general, for *organization* entity types we use the relation *based-in0* to connect it with *gpe0*. However, we define additional relations that reflect the hierarchical entity type chain: *agency-of* to connect *gov org*, *ministry-of* to connect *ministry* with *location* sub-types. This relation type hierarchy coupling with named entity *type* tag hierarchy is not strict and we only define the corresponding relation types if (i) it results in meaningful and interpretable relation, and (ii) if the dataset contains actual instances of a particular relation type. Similarly to entity type annotations, this approach allows to quickly extend the annotations to more granular types by filtering on a particular generic (parent) relation type and dividing into further sub-types if they are backed up by enough instances in the dataset.

Furthermore, unlike the current mainstream relation datasets [21, 38, 52, 58, 64], we thus also allow relations between the same two entities in our corpus to have multiple labels. Thus, in the example in Fig. 1 we note that the relation between *Harry* and *Britain* is both *royalty-of* and *citizen-of*. Additionally, we define further logical rules to automatically guarantee the consistency of the relations and their types. For instance, in the example of Fig. 1, the *parent-of* relation between “William” and “Charles” entities can be derived from the fact that “Harry” is a *sibling-of* “Charles” and “William” is *parent-of* “Harry”. This way, we also introduce an additional reasoning layer to predict the relations in our dataset. A complete list of logical rules is defined in Appendix E. Note that this list also includes rules related to the hierarchy of relations. This gives us flexibility to modify the hierarchical structure of our annotation schema.

3.3 Inter-Annotator Agreement and Annotation Refinement

In order to assess and further improve the quality of our dataset we annotated 100 randomly selected news articles (12.5% of our dataset) from scratch. This work is done by a second independent expert annotator. The annotations in this pass are performed by following the already defined annotation schema based on the annotation process in *exploratory* and *schema-driven* passes. We use this second annotated subset to measure the inter-annotator agreement and subsequently determine the parts of the dataset that still need to be improved. In order to measure the agreement we use the Cohen’s kappa coefficient [11] whose general formula is described in Eq. (1). Where p_o represents the observed agreement between the two annotators and p_e is the expected



agreement between the annotators (i.e., agreement by chance).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

More specifically, in our case we calculate the observed probability p_o as in Eq. (2) where N is the number of annotated items and $agree(arg1, arg2)$ returns 1 if $arg1$ is equal to $arg2$ and 0 otherwise. $A_{i,j}$ on the other hand, is the annotation made by annotator i for item j . Thus, p_o can be interpreted as the fraction of the labels two annotators agree, also called *percent agreement* [41, 50].

$$p_o = \frac{\sum_{j=1}^N agree(A_{1,j}, A_{2,j})}{N} \tag{2}$$

To calculate the expected agreement probability we use the formulation in Eq. (3) which can be interpreted as an estimation of the probability that both annotator 1 and 2 will annotate a particular item with label l . In this context, $n_{i,l}$ is the number of items the annotator i annotated with label l and L is the total number of labels.

$$p_e = \sum_{l=1}^L \frac{n_{1,l}}{N} \cdot \frac{n_{2,l}}{N} = \frac{1}{N} \sum_{l=1}^L n_{1,l} n_{2,l} \tag{3}$$

For multi-label annotations where it is possible to assign multiple classes for a particular annotation item (i.e., named entity and relation types), we report a weighted kappa score. Table 5 presents a summary of Cohen’s kappa scores for each of the multi-task layers before and after the dataset revision performed in this pass. We observe that even before the revision the lowest kappa scores for relation annotation of 0.6594 indicate a moderate to substantial agreement [31, 41]. We further improve this score to 0.8729 which is considered strong [41] to almost perfect [31] agreement. We achieved this by carefully examining and analyzing the discrepancies, fixing the dataset and adapting accordingly the annotation guidelines. It is also worth noting that the agreement score for *relation detection* is always slightly lower than for *relation classification*. This is because manually detecting relations on document level is rather hard and time-consuming task that requires a careful reading and interpretation of the document content. However, once the relation has been identified, it is rather straightforward to assign the correct type given the clearly defined annotation guidelines. Furthermore, we observe an opposite phenomenon for named entities where, despite the fact that we achieve almost perfect agreement on both entity detection and classification, the agreement for detection is slightly higher (0.9673) than for classification (0.9026). The explanation for this is that named entity detection is rather straightforward and most of the entities can be identified by a single document scan. However, the classification process, even with defined guidelines, is not always clear cut and aggravated by a substantially higher number of entity tags (331) in comparison to the number of relation types (69).



Task	Before Revision	After Revision
Named Entity	0.8497	0.8703
Named Entity Detection	0.9665	0.9673
Named Entity Classification	0.8812	0.9026
Coreference	0.9302	0.9324
Entity Linking	0.9280	0.9320
Relation	0.6594	0.8729
Relation Detection	0.7686	0.8727
Relation Classification	0.8118	0.9666

Table 5: The inter-annotation agreement scores between all the different annotation tasks *before* and *after* the dataset revision performed after analyzing the discrepancies between the parallel annotations. The scores are reported using Cohen’s kappa or weighted Cohen’s kappa metric depending on whether the nature of annotations is single (NER Detection, Keyword Detection, Entity Linking, Relation Detection) or multi-label (NER Classification and Relation Classification) respectively.

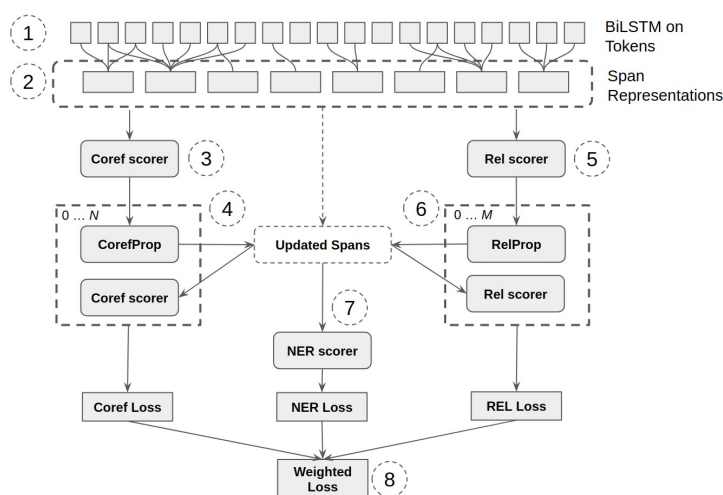


Fig. 4: DyGIE architecture, the numbers in the circles represent the order of the execution of different components.

4.1 Baseline Architecture

We use a recently proposed span-based DyGIE [39] architecture in order to demonstrate how the different annotated layers of DW-Articles dataset can be used jointly to boost the performance of each of the individual tasks. Figure 4 illustrates the architecture of DyGIE. The input to the model consists of a concatenation of GLoVe [44] and character embeddings. A further layer of bidirectional LSTM is applied. Each of the span representations is obtained by concatenating the hidden LSTM representations at the begin and end span



positions. We denote these original span representations as $G^0 = \{\mathbf{g}^0_1, \dots, \mathbf{g}^0_T\} : \mathbf{g}^0_i \in \mathbb{R}^d$ that are associated to each of the $S = s_1, \dots, s_T$ within-document textual spans where T is the total number of spans in the document. This contrasts with sentence-based original DyGIE implementation provided in [39, 57]. The DyGIE architecture provides the option to further iteratively refine these span representations by using graph propagation techniques on coreference (N propagation iterations) after which an updated span representation $G^N = \{\mathbf{g}^N_1, \dots, \mathbf{g}^N_T : \mathbf{g}^N_i \in \mathbb{R}^d\}$ is obtained to be used as input to relation propagation and scorer modules. Similarly, the relation graph propagation module is executed (M propagation iterations). This propagation yields updated span representation $G^{N+M} = \{\mathbf{g}^{N+M}_1, \dots, \mathbf{g}^{N+M}_T : \mathbf{g}^{N+M}_i \in \mathbb{R}^d\}$

that is used as input to named entity scorer model. Accordingly, we expect to demonstrate that by using these graph propagation techniques, we can boost the performance of our individual relation and NER tasks. Below, we provide more in detail formulations of DyGIE model used in our experiments. We follow the notation proposed originally in [39]. Furthermore, we explain the adaptations that we had to make in order to apply it to DW-Articles dataset. **Coreference Scoring and Propagation** — The coreference confidence scores between span i and j for propagation iteration t denoted as $P^t(i, j)$ is calculated as in Eq. (4),

$$P_C^t(i, j) = \frac{\exp(V^t(i, j))}{\sum_{j \in B_C(i)} \exp(V_C^t(i, j))} \quad (4)$$

Where $V_C^t(i, j)$ denotes a Feed-forward Neural Network (FFNN) that takes as input a following concatenation of span representations $[\mathbf{g}^t_i, \mathbf{g}^t_j, \mathbf{g}^t_i \otimes \mathbf{g}^t_j]$ and outputs a scalar score. The $B_C(i)$ are the antecedent spans of s_i upper-bounded by the beam size b_c , which is highly dataset-dependent hyper-parameter, and in our experiments we set it to 125, which is enough to cover all of the document-level co-references in DW-Articles dataset. In case $N > 0$ indicating that we want to calculate coreference propagation, the span update vector $\mathbf{u}_C^t \in \mathbb{R}^d$ is first calculated as indicated in Eq. (5) to be later used in Eq. (7) to update span representations. Essentially, \mathbf{u}_C^t consists in a weighted sum of neighbouring span representations \mathbf{g}_j^t where the weights are the coreference confidence scores $P_C^t(i, j)$ calculated in Eq. (4).

$$\mathbf{u}_C^t(i) = \sum_{j \in B_C(i)} P_C^t(i, j) \mathbf{g}_j^t \quad (5)$$

Relation Scoring and Propagation — Unlike a single coreference scalar score $V_C^t(i, j)$, for relation a vector of length L_R is used to denote the possible relation scores between two spans. The complete relation scoring tensor is denoted as $V_R^t \in \mathbb{R}^{b_R \times b_R \times L_R}$ where b_R is the number of spans involved in relation, and L_R is the number of possible relation classes. To calculate the optimal b_R for a particular document, we use a similar approach as the one adopted in the implementation in [57], namely to set b_R as a proportion of number of tokens. The only difference is that while in [57] the proportion is taken with



respect to the number of tokens in a sentence, we take it with respect to the number of tokens in a document. We use a proportion of 0.20 which covers more than 98% of relation mention pairs in the DW-Articles dataset. Similarly as with coreference, in order to calculate the L_R -length vector of relation scores $\mathbf{V}_R^t(i, j)$, a FFNN is used with $[\mathbf{g}_i^t, \mathbf{g}_j^t, \mathbf{g}_i^t \mathcal{B} \mathbf{g}_j^t]$ as input. Unlike the original DyGIE version that uses a maximum predicted score $\mathbf{V}^t(i, j)$ between the span i and j to predict a single relation type, in DW-Articles we use a threshold on $\mathbf{V}_R^t(i, j)$ after the final relation propagation to predict a set of possible relation types. This is because, as shown in Table 4, an important number of relations in DW-Articles dataset between two mentions (more than 30% of mention pairs) is multi-label. Similarly as with coreference (Eq. (5)), a relation span update vector is calculated as described in Eq. (6),

$$\mathbf{u}_i^t = \sum_{j \in B_R} f(\mathbf{V}_R^t(i, j)) \mathbf{A}_R \mathcal{B} \mathbf{g}_j^t \quad (6)$$

Where B_R is the beam of top b_R spans ranked by a relation pruner (FFNN), $\mathbf{A}_R \in \mathbb{R}^{L_R \times d}$ is a trainable projection tensor, and f is a non-linearity function (ReLU).

Updated Span Representation — The coreference (C) and relation (R) update vectors defined in Eq. (5) and (6) respectively are used to update the span representation for the next iteration $t+1, \dots, N+M$ as described formally in Eq. (7). First, the N coreference iterations ($x=C$) are executed, followed by M relation iterations ($x=R$),

$$\mathbf{f}_x^t(i) = g(\mathbf{W}_x^t [\mathbf{g}_i^t, \mathbf{u}_i^t(i)]) \quad (7)$$

$$\mathbf{g}_i^{t+1} = \mathbf{f}_x^t(i) \mathcal{B} \mathbf{g}_i^t + (1 - \mathbf{f}_x^t(i)) \mathcal{B} \mathbf{u}_i^t(i)$$

Where $x \in \{C, R\}$, $\mathbf{W}_x^t \in \mathbb{R}^{d \times 2d}$, g is a sigmoid function, $\mathbf{f}_x^t \in \mathbb{R}^d$ can be interpreted as a gating vector that acts as a switch between the current span representations \mathbf{g}_i^t , and the update span vector $\mathbf{u}_i^t(i)$.

Entity scoring — The mention scoring is done on the final G^{N+M} span representations using FFNN. Unlike the original DyGIE implementation that predicts single label for a particular span, we use a multi-label approach, since each of the entities in DW-Articles dataset can be assigned multiple tag labels (see Table 3).

Loss Function — the total loss in DyGIE is a weighted average of individual losses of coreference, relation and entity scoring.

4.2 Decoding and Prediction

In order to use a DyGIE model to make predictions on the DW-Articles dataset the individual predictions of clusters, relations and entity mentions have to be integrated. First, K coreference clusters are predicted, we denote them with $C^p = \{c_1^p, \dots, c_K^p\}$. Each of the predicted clusters c_j^p is composed by one or more spans. Conceptually, these predicted clusters denote the entities with the



corresponding mentions. In order to get the predicted entity tags, we recur to the individual predicted span tags by *NER scorer* $T^p = \{t_1^p, \dots, t_7^p\}$ for each p of the T originally extracted tags. Each of the t_1^p is composed by zero or more predicted entity tags. The resulting tags assigned to each of the predicted entity clusters c_i^p are composed by the union of the tags predicted for each of the span mentions $s_j \in c_i^p$. Similarly, in order to predict relations between two clusters c_I and c_J , we group (using union operation) all the predicted multi-label relation types between individual spans $s_i \in c_I$ and $s_j \in c_J$ derived from the relation scores $\mathbf{V}^t(i, j, R)$.

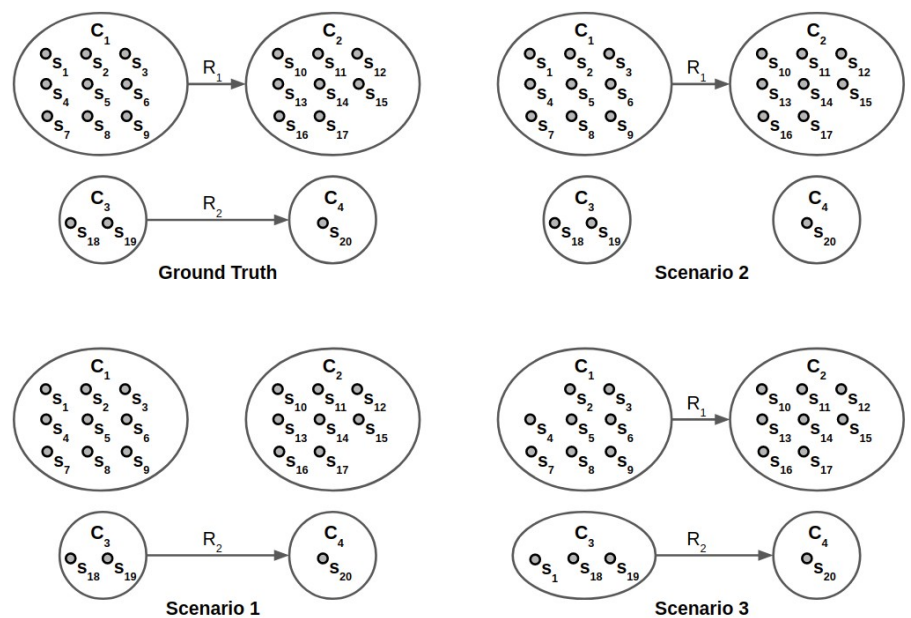


Fig. 5: Relation metrics motivation.

We use traditional MUC [56], B-CUBED [4] and CEAF [40] metrics to evaluate the coreference in our models. Similarly as it is done in other related work [9, 10, 13, 34, 35, 60], we report the average of these three metrics.

For NER score, we calculate the precision, recall and F1 scores on mention level (M-F1) as well as on soft-entity (S-F1) weighted level. While, M-F1 measure provides a good way to evaluate the impact of our model on individual mention-level, it can get biased towards entities with big number of mentions, this is why S-F1 was developed .



In order to report on relations, we contemplate mention level (M-F1), soft-entity level (S-F1), as well as hard cluster level (H-F1) metrics. We consider S-F1 as our main metric to evaluate the performance of the models since we consider it captures the balance between mention-level and entity-level relation scoring. In order to illustrate the advantage of using S-F1 metric we recur to examples illustrated in Fig. 5. Consider two relations R_1 and R_2 that involve four clusters $\{C_1, C_2, C_3, C_4\}$ each one composed by mentions represented as text spans $\{s_1, \dots, s_{20}\}$. The first one (R_1) is between two relatively big entity clusters that are represented as C_1 and C_2 where $|C_1| = 9$ and $|C_2| = 8$. The relation R_2 , on the other hand, is between two small clusters C_3 and C_4 where $|C_3| = 2$ and $|C_4| = 1$. We follow the general equations for calculation of Precision (Pr), Recall (Re) and F_1 scores described in Equations (8) - (10) in order to build the Table 6 where we compare mention-based (M-Pr, M-Re and M-F1), soft entity-based (S-Pr, S-Re and S-F1), and hard entity-based (H-Pr, H-Re and H-F1) metrics for all three scenarios described in Figure 5.

$$Pr = \frac{tp}{tp + fn} \quad (8)$$

$$Re = \frac{tp}{tp + fn} \quad (9)$$

$$F_1 = 2 \cdot \frac{Pr \cdot Re}{Pr + Re} \quad (10)$$

The disadvantage of using purely mention-based (M-F1) relation metric as in previous work [5, 36, 39, 51, 62, 63, 64], is that the entity relations R_1 and R_2 are assigned different weight. More specifically, M-F1 is highly influenced by the relation R_1 between two big clusters. Thus, in case of *Scenario 1* in Figure 5 where the model fails to predict R_1 relation between C_1 and C_2 , it transforms in a huge drop in M-F1. This is because there are $|C_1| \cdot |C_2| = 8 \cdot 9 = 72$ mention-level relations between C_1 and C_2 . This results in 72 false negative (fn) cases compared to only $|C_3| \cdot |C_4| = 1 \cdot 2 = 2$ true positive (tp) cases given by correct prediction of R_2 . This translates in a very low recall (0.027), which produces a dis-proportionally low F1 score (0.053). Conversely, *Scenario 2* presents an opposite situation where the relation R_1 between entity clusters C_1 and C_2 is predicted correctly, but not the relation R_2 between the cluster C_3 and C_4 . The resulting mention-level F1 (M-F1 of 0.986) is driven by high recall that is caused by a big number of mentions connected by R_1 compared to R_2 . In reality, however, only 1 out of 2 relations between entities is correctly predicted. In both cases, this problem can be solved by considering what we denominate *Hard-Entity Level* scores where both, the entity cluster content as well as predicted relation between entity clusters have to be correctly predicted in order to be counted as true positives. However, this metric seems to be overly strict in situations such as the one depicted in *Scenario 3*. Here, only one mention given by span s_1 was assigned to the wrong entity cluster C_3 instead of C_1 , which makes hard entity match fail, assigning 0.0 in H-Pr, H-Re, and H-F1 scores.



	Mention Level			Soft-Entity Level			Hard-Entity Level		
	M-Pr	M-Re	M-F1	S-Pr	S-Re	S-F1	H-Pr	H-Re	H-F1
Ground Truth	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Scenario 1	1.000	0.027	0.053	1.000	0.500	0.667	1.000	0.500	0.667
Scenario 2	1.000	0.973	0.986	1.000	0.500	0.667	1.000	0.500	0.667
Scenario 3	0.985	0.892	0.936	0.833	0.944	0.885	0.000	0.000	0.000

Table 6: Comparison of different relation metrics.

As a consequence, we propose a *Soft-Entity Level* metric that is a relaxed version of *Hard-Entity Level*. We do this by taking into account the number of correctly connected mention spans (mention-based level metric), but re-weighting by the connected cluster sizes in order to not let the metric be dominated by clusters with large number of entity mentions. We do this by defining G_{rel}^C and P_{rel}^C as a set of ground truth and predicted relation tuples on a cluster level such that $G_{rel}^C \in \{(C_s, p, C_o), \dots, (C_{R_g^C}, p_{R_g^C}, C_{R_g^C})\}$ and $P_{rel}^C \in \{(C_s, p, C_o), \dots, (C_{R_p^C}, p_{R_p^C}, C_{R_p^C})\}$ where R_g^C and R_p^C are the total numbers of cluster-level ground truth and predicted relations respectively. Similarly, we define G_{rel}^M and P_{rel}^M to refer to the set of expanded ground truth and predicted mention-level relation tuples. We further define a function $M(C_s, p, C_o)$ that, given a subject (C_s) and object (C_o) relation entity clusters as well as a relation predicate p , returns a set of mention-level relation tuples between spans of C_s and C_o of type p where $M(C_s, p, C_o) \subseteq C_s \times C_o$. Next, from the equations (8)-(10), we know that in order to calculate precision, recall and F1 metrics, we first need to compute the number of true positives (tp), false positives (fp) and false negatives (fn). This is done, in equations (11)-(14). As it can be noted, we define two true positives metrics. The first one (tp_p) is re-weighted by the number of mention-level relations between predicted clusters, while the second one (tp_g) is re-weighted by the number of mention-level relations between gold clusters. This is needed to be able to achieve the constraints described in equations (15) and (16), in order for the precision and recall scores always be between 0 and 1.

$$tp_p = \frac{|M(C_s, p, C_o) \cap G_{rel}^M|}{|M(C_s, p, C_o)|} \quad (11)$$

$$tp_g = \frac{|M(C_s, p, C_o) \cap P_{rel}^M|}{|P_{rel}^M|} \quad (12)$$

$$fp = 1.0 - \frac{|M(C_s, p, C_o) \cap P_{rel}^M|}{|M(C_s, p, C_o)|} \quad (13)$$

$$fn = 1.0 - \frac{|M(C_s, p, C_o) \cap G_{rel}^M|}{|G_{rel}^M|} \quad (14)$$



$$tp.p + fp = \# \text{ of predicted rels} \quad (15)$$

$$tp.g + fn = \# \text{ of correct rels} \quad (16)$$

Equations (17)-(19), on the other hand, describe how precision(E-Pr), recall(E-Re) and F1(E-F1) scores of the Soft-Entity Metric are calculated.

$$E-Pr = \frac{tp.p}{tp.p + fp} \quad (17)$$

$$E-Re = \frac{tp.g}{tp.g + fn} \quad (18)$$

$$E-F1 = 2 \cdot \frac{E-Pr \cdot E-Re}{E-Pr + E-Re} \quad (19)$$

Model Setup	Coreference			Relation		Tag
	AVG-F1	M-F1	E-F1	M-F1	E-F1	
Independent NER	-	-	-	86.3	?	
Independent Rel	88.7	?	68.9	-	-	
Joint	89.5	?	71.5	87.5	?	
Joint+BERT	?	?	?	?	?	

Table 7: Main Results. The missing results (? marks) will be completed in the final version of the manuscript.

We experiment by training DyGIE model on three tasks: NER, coreference, and relation extraction. Table 7 describes the main results of training on independent task models (*Independent NER* and *Independent Rel*), compared to a joint approach where all the tasks are trained jointly. As it can be observed, the joint setup has a significantly better performance than individual tasks. This demonstrates the benefit of our dataset in multi-task setting. Furthermore, we perform additional experiments in order to study how CorefProp and RelProp iterations affect the performance of individual tasks. Thus, in Table 8 we can observe that CorefProp has a positive effect of 1.7 F1 points on our M-F1 NER with respect to *Independent NER* model. The RelProp, on the other hand, also produces an improvement of 0.4 F1 points. Similarly, we observe a boost of performance on the *Independent Rel* model of 3 F1 points when trained jointly with CorefProp.

Additionally, we explore the effect of the number of propagation (CorefProp and RelProp) iterations on the final F1 score on the NER and Relation extraction tasks, which is illustrated in Figure 6. We observe that the main benefit comes from CorefProp module both for NER as well as for Relation prediction modules. The improvement coming from RelProp on NER module is less significative. Furthermore, we still note that potentially there is still



Model Setup	NER	
	M-F1	E-F1
Independent NER	85.6	?
+CorefProp	87.3	?
+RelProp	86.1	?
+BothProp	?	?
Independent NER+BERT	?	?
+CorefProp	?	?
+RelProp	?	?
+BothProp	?	?

Table 8: Impact of CorefProp and RelProp on NER. The missing results (? marks) will be completed in the final version of the manuscript.

Model Setup	Relation	
	M-F1	E-F1
Independent Rel	?	68.87
+CorefProp	?	72.09
+RelProp	?	?
+BothProp	?	?
Independent Rel+BERT	?	?
+CorefProp	?	?
+RelProp	?	?
+BothProp	?	?

Table 9: Impact of CorefProp and RelProp on Relation. The missing results (? marks) will be completed in the final version of the manuscript.

space for improvement by adding more CorefProp (N) as well as RelProp (M) Iterations. Strangely enough, we also note an initial drop in F1 performance score for $N = 1$ when CorefProp is applied to relation prediction module. We are going to study the cause of this phenomena in the final version of the paper.

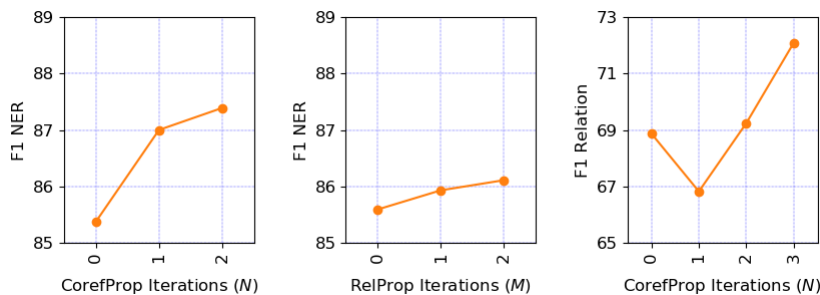


Fig. 6: Impact of the Coref and Relation Propagation Iterations different tasks.



In this work we introduce DW-Articles, a manually annotated multi-task dataset that comprises NER, Coreference, Relation Extraction and Entity Linking as main tasks. We show how DW-Articles is different from the mainstream datasets by focusing on document-level and entity-centered annotations. This makes also the predictions on this dataset more challenging by having not only to consider explicit, but also implicit document-level interactions between entities. Furthermore, we show how different tasks inter-relate with each other by training joint models and demonstrating their superior performance compared to single-task approaches.

Acknowledgements Part of the research leading to these results has received funding from (i) the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 761488 for the CPN project, and (ii) the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

REFERENCES

1. Akbik, A., Bergmann, T., Vollgraf, R.: Pooled contextualized embeddings for named entity recognition. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 724–728 (2019)
2. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 546–555 (2017)
3. Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., Auli, M.: Cloze-driven pretraining of self-attention networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5363–5372 (2019)
4. Bagga, A., Baldwin, B.: Algorithms for scoring coreference chains. In: The first international conference on language resources and evaluation workshop on linguistics coreference, vol. 1, pp. 563–566. Citeseer (1998)
5. Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: Adversarial training for multi-context joint entity and relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2830–2836 (2018)
6. Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications* **114**, 34–45 (2018)
7. Blanco, R., Ottaviano, G., Meij, E.: Fast and space-efficient entity linking for queries. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 179–188. ACM (2015)



8. Chiu, J.P., Nichols, E.: Transactions of the Association for Computational Linguistics **4**, 357–370 (2016)
9. Clark, K., Manning, C.D.: Entity-centric coreference resolution with model stacking. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1405–1415 (2015)
10. Clark, K., Manning, C.D.: Improving coreference resolution by learning entity-level distributed representations. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 643–653 (2016)
11. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
12. Derczynski, L., Nichols, E., van Erp, M., Limsopatham, N.: Results of the wnut2017 shared task on novel and emerging entity recognition. In: Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 140–147 (2017)
13. Durrett, G., Klein, D.: Easy victories and uphill battles in coreference resolution. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1971–1982 (2013)
14. Ellis, J., Getman, J., Fore, D., Kuster, N., Song, Z., Bies, A., Strassel, S.M.: Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In: TAC (2015)
15. Ellis, J., Getman, J., Strassel, S.M.: Overview of linguistic resources for the tac kbp 2014 evaluations: Planning, execution, and results. In: Proceedings of TAC KBP 2014 Workshop, National Institute of Standards and Technology, pp. 17–18 (2014)
16. Eshel, Y., Cohen, N., Radinsky, K., Markovitch, S., Yamada, I., Levy, O.: Named entity disambiguation for noisy text. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 58–68 (2017)
17. Francis-Landau, M., Durrett, G., Klein, D.: Capturing semantic similarity for entity linking with convolutional neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1256–1261 (2016)
18. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics (1996)
19. Han, X., Sun, L.: An entity-topic model for entity linking. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 105–115 (2012)
20. Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., Sun, M.: Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In: Proceedings of the 2018 Conference on Empirical



- Methods in Natural Language Processing, pp. 4803–4809 (2018)
21. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Ó S éaghdha, D., Pad ó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pp. 94–99. Association for Computational Linguistics (2009)
 22. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 782–792. Association for Computational Linguistics (2011)
 23. Ji, H., Grishman, R., Dang, H.T., Griffitt, K., Ellis, J.: Overview of the tac 2010 knowledge base population track. In: Third text analysis conference (TAC 2010), vol. 3, pp. 3–3 (2010)
 24. Ji, H., Nothman, J., Hachey, B., Florian, R.: Overview of tac-kbp2015 tri-lingual entity discovery and linking. In: TAC (2015)
 25. Ji, H., Pan, X., Zhang, B., Nothman, J., Mayfield, J., McNamee, P., Costello, C., Hub, S.I.: Overview of tac-kbp2017 13 languages entity discovery and linking. In: TAC (2017)
 26. Jia, R., Wong, C., Poon, H.: Document-level n-ary relation extraction with multiscale representation learning. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3693–3704 (2019)
 27. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: Genia corpora semantically annotated corpus for bio-textmining. *Bioinformatics* **19**(suppl 1), i180–i182 (2003)
 28. Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. *CoNLL 2018* p. 519 (2018)
 29. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: Collective annotation of wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 457–466 (2009)
 30. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of NAACL-HLT, pp. 260–270 (2016)
 31. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *biometrics* pp. 159–174 (1977)
 32. Le, P., Titov, I.: Improving entity linking by modeling latent relations between mentions. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1595–1604 (2018)
 33. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics* **39**(4), 885–916 (2013)



34. Lee, K., He, L., Lewis, M., Zettlemoyer, L.: End-to-end neural coreference resolution. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 188–197 (2017)
35. Lee, K., He, L., Zettlemoyer, L.: Higher-order coreference resolution with coarse-to-fine inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 687–692 (2018)
36. Li, Q., Ji, H.: Incremental joint extraction of entity mentions and relations. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 402–412 (2014)
37. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2124–2133 (2016)
38. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3219–3232 (2018)
39. Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., Hajishirzi, H.: A general framework for information extraction using dynamic span graphs. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3036–3046 (2019)
40. Luo, X.: On coreference resolution performance metrics. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, pp. 25–32. Association for Computational Linguistics (2005)
41. McHugh, M.L.: Interrater reliability: the kappa statistic. *Biochemia medica*: *Biochemia medica* **22**(3), 276–282 (2012)
42. Medero, J., Strassel, S.: Readme ace 2005. <https://catalog ldc.upenn.edu/docs/LDC2006T06/README> (2006). [Online; accessed 14-March-2020]
43. Peng, N., Poon, H., Quirk, C., Toutanova, K., Yih, W.t.: Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics* **5**, 101–115 (2017)
44. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014)
45. Quirk, C., Poon, H.: Distant supervision for relation extraction beyond the sentence boundary. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 1171–1182 (2017)
46. Rau, L., Krupka, G., Jacobs, P.: Ge ntoolset: Muc-4 test results and analysis. In: FOURTH MESSAGE UNDERSTANDING CONFERENCE (MUC-4), Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992 (1992)



47. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 148–163. Springer (2010)
48. Riedel, S., Yao, L., McCallum, A., Marlin, B.M.: Relation extraction with matrix factorization and universal schemas. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 74–84 (2013)
49. Sang, E.F.T.K., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142–147 (2003)
50. Scott, W.A.: Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly* pp. 321–325 (1955)
51. Soares, L.B., FitzGerald, N., Ling, J., Kwiatkowski, T.: Matching the blanks: Distributional similarity for relation learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2895–2905 (2019)
52. Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., Ma, X.: From light to rich ere: annotation of entities, relations, and events. In: Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pp. 89–98 (2015)
53. Strassel, S.M., Przybocski, M.A., Peterson, K., Song, Z., Maeda, K.: Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In: LREC (2008)
54. Sun, C., Gong, Y., Wu, Y., Gong, M., Jiang, D., Lan, M., Sun, S., Duan, N.: Joint type inference on entities and relations via graph convolutional networks. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1361–1370 (2019)
55. Verga, P., Strubell, E., McCallum, A.: Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 872–884 (2018)
56. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the 6th conference on Message understanding, pp. 45–52. Association for Computational Linguistics (1995)
57. Wadden, D., Wennberg, U., Luan, Y., Hajishirzi, H.: Entity, relation, and event extraction with contextualized span representations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5788–5793 (2019)
58. Walker, C., Strassel, S., Medero, J., Maeda, K.: Ace 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia **57** (2006)



59. Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al.: Ontonotes release 5.0 ldc2013t19. Linguistic Data Consortium, Philadelphia, PA **23** (2013)
60. Wiseman, S., Rush, A.M., Shieber, S.M.: Learning global features for coreference resolution. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 994–1004 (2016)
61. Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., Sun, M.: Docred: A large-scale document-level relation extraction dataset. arXiv preprint arXiv:1906.06127 (2019)
62. Zhang, M., Zhang, Y., Fu, G.: End-to-end neural relation extraction with global optimization. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1730–1740 (2017)
63. Zhang, Y., Qi, P., Manning, C.D.: Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2205–2215 (2018)
64. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 35–45 (2017)



Appendices

A HIERARCHICAL ENTITY TYPE STATISTICS

Entity Type	# Entities	%Entities	# Mentions	%Mentions
ENTITY	13152	56.9 %	30716	70.8 %
location	4956	21.4%	11547	26.6%
gpe	3965	17.1%	9830	22.7%
gpe0	2225	9.6%	6559	15.1%
gpe2	1497	6.5%	2873	6.6%
gpe1	244	1.1%	406	0.9%
regio	479	2.1%	916	2.1%
facility	258	1.1%	384	0.9%
waterbody	91	0.4%	145	0.3%
organization	3419	14.8%	8128	18.7%
media	659	2.8%	984	2.3%
igo	547	2.4%	1992	4.6%
so	171	0.7%	912	2.1%
party	381	1.6%	949	2.2%
company	368	1.6%	932	2.1%
sport_team	367	1.6%	1106	2.5%
governmental organisation	342	1.5%	636	1.5%
agency	228	1.0%	444	1.0%
ministry	81	0.4%	104	0.2%
armed_movement	107	0.5%	373	0.9%
ngo	101	0.4%	172	0.4%
todo_org	98	0.4%	142	0.3%
education_org	79	0.3%	98	0.2%
research_center	55	0.2%	80	0.2%
policy_institute	54	0.2%	85	0.2%
person	3396	14.7%	8259	19.0%
politician	1183	5.1%	3324	7.7%
head_of_state	379	1.6%	1270	2.9%
head_of_gov	247	1.1%	673	1.6%
minister	217	0.9%	458	1.1%
sport_player	405	1.8%	844	1.9%
todo_per	322	1.4%	666	1.5%
artist	262	1.1%	586	1.4%
writer	64	0.3%	210	0.5%
actor	56	0.2%	122	0.3%
filmmaker	56	0.2%	82	0.2%
politics_per	209	0.9%	457	1.1%
manager	104	0.4%	297	0.7%
journalist	104	0.4%	153	0.4%
gov_per	98	0.4%	160	0.4%
offender	75	0.3%	347	0.8%
activist	68	0.3%	181	0.4%
clergy	59	0.3%	137	0.3%
sport_coach	57	0.2%	129	0.3%
misc	832	3.6%	1679	3.9%
work_of_art	173	0.7%	246	0.6%
film_title	79	0.3%	119	0.3%
award	88	0.4%	196	0.5%
treaty	81	0.4%	150	0.3%
product	65	0.3%	148	0.3%
todo_misc	59	0.3%	99	0.2%
event	354	1.5%	701	1.6%
sport_competition	183	0.8%	410	0.9%
war	56	0.2%	74	0.2%
language	90	0.4%	165	0.4%
ethnicity	84	0.4%	242	0.6%
VALUE	5903	25.5 %	7104	16.4 %
time	2907	12.6%	3608	8.3%
role	2390	10.3%	2865	6.6%
money	606	2.6%	631	1.5%
OTHER	2735	11.8 %	5493	12.7 %
loc0-x	1596	6.9%	3827	8.8%
footer	413	1.8%	413	1.0%
loc-x	353	1.5%	585	1.3%
religion-x	235	1.0%	486	1.1%
none	78	0.3%	97	0.2%
skip	75	0.3%	93	0.2%



DW-Articles: A multi-task document-level entity-driven dataset for				31
<i>EXPERT</i>	9	0.0%	26	0.1%
<i>ALIAS</i>	7	0.0%	43	0.1%
TOTAL	23133	100.0%	43383	100.0%

Table 10: Entity type statistics...



B INTERACTION OF ENTITY TOPIC AND TYPE TAGS

The content of each of the cells consists of the *Entity* sub-types of different hierarchy levels represented by the column that belong to the topic represented by the row.

	person	organization	event	location	misc
politics	head_of_gov, head_of_state, minister, politician regional, politician local, politician national, politics_candidate, politician, politics_per, activist, gov_per	politics_institution, politics_org, party, ngo, igo, so, policy_institute, movement, agency, ministry, military_alliance	summit_meeting, scandal, politics_event	politics_facility	politics_misc, politics_project, treaty, report
culture	character, culture_per, artist, writer, actor, filmmaker, musician	music_band, culture_org	festival, filmfestival	culture_facility	art_title, culture_title, exhibition_title, culture_misc, work_of_art, book_title, film_title, tv_title, music_title, theatre_title, musical_title, film_award, book_award, music_award, tv_award, column_title
education	education_teacher, education_per, education_student	education_org		education_facility	education_study
religion	religion_deity, religion_clergy	religion_org	religion_event	religion_facility	religion, religion_misc
human	royalty				film_award, book_award, award, music_award, tv_award, sport_award
conflict	military_personnel, military_rebel	army, military_alliance, armed_movement	war, protest	military_facility	military_equipment, military_mission
media	journalist	media			
science	researcher, science_per	research_center			species, research_journal, technology
sport	sport_player, sport_coach, sport_head, sport_referee, sport_per	sport_team, sport_org	sport_competition	sport_facility	sport_award
labour	union_head, union_member, union_rep	union			
business	manager, employee, business_per	company, business_org, trade_fair, market_exchange, business_advocacy		business_facility	product, brand, market_index, business_misc
health	health_per	health_org		health_facility	health_disease, health_drug
justice	offender, advisor, victim, judge, police_per, justice_per	court, criminal_org, police_org, justice_org		prison	justice_misc, case
weather			weather_storm		

Table 11: Structural illustration of the NER entity types in DW-Articles dataset. The columns represent the main entity types (person, organization, event, location, misc). The rows represent the different content topics in which we categorize the named entities. Each of the cells contains the possible NER subtypes (of different hierarchy levels)





C ENTITY LINKING STATISTICS

Table 12 describes the statistics of linked entities with respect to the total number of entities in each of the *Entity* sub-types. The columns *% Linked Entities* and *% Linked Mentions* indicate the percentage of annotated linked entities and mentions with respect to the total number of annotated entities/mentions in a particular *Entity Type* category. The column *Prior Linking* on the other hand, indicates the accuracy on test set of linking an entity to the most frequent link used in training set. For example, we observe that geopolitical entities *gpe* (e.g., countries, cities, states, etc.) have very high *prior linking* accuracy. This indicates that there is high repetition of geopolitical names between the annotated articles. Other entity types such as *facility* or *person* have lower *prior linking* scores, indicating that there is higher variation of these entities in our dataset and the automatic linking process can be potentially more challenging on these entity types. The entity linking annotation is carried out by the expert annotator.

Entity Type	# Linked Entities	% Linked Entities	# Linked Mentions	% Linked Mentions	Prior Linking
<i>ENTITY</i>	11741	89.3%	28138	91.6%	71.4%
location	4862	98.1%	11425	98.9%	85.9%
gpe	3938	99.3%	9800	99.7%	89.8%
regio	456	95.2%	885	96.6%	83.8%
facility	228	88.4%	350	91.1%	24.6%
waterbody	90	98.9%	144	99.3%	83.3%
district	37	94.9%	42	93.3%	33.3%
organization	3132	91.6%	7700	94.7%	70.4%
media	622	94.4%	936	95.1%	80.5%
igo	525	96.0%	1943	97.5%	76.4%
party	358	94.0%	897	94.5%	76.2%
company	320	87.0%	858	92.1%	71.4%
sport_team	366	99.7%	1105	99.9%	71.0%
person	2629	77.4%	6619	80.1%	56.4%
politician	1162	98.2%	3296	99.2%	66.4%
sport_player	404	99.8%	843	99.9%	34.4%
todo_per	126	39.1%	259	38.9%	67.9%
artist	246	93.9%	531	90.6%	5.6%
politics_per	126	60.3%	277	60.6%	54.7%
misc	616	74.0%	1356	80.8%	58.0%
work_of_art	141	81.5%	205	83.3%	9.1%
award	72	81.8%	174	88.8%	63.6%
treaty	60	74.1%	115	76.7%	71.4%
product	50	76.9%	126	85.1%	53.8%
todo_misc	35	59.3%	73	73.7%	40.0%
event	320	90.4%	656	93.6%	50.0%
sport_competition	163	89.1%	385	93.9%	62.0%
war	56	100.0%	74	100.0%	NaN
summit_meeting	15	68.2%	27	67.5%	100.0%
holiday	21	95.5%	39	97.5%	100.0%
history	17	89.5%	28	93.3%	100.0%
TOTAL	11741	89.3%	28138	91.6%	71.4%

Table 12: Entity linking statistics.



D RELATION TYPE STATISTICS

Relation Type	Related Ent. Pairs	%Related Ent. Pairs	# Related Men. Pairs	% Related Men. Pairs
based. in0	2359	14.150%	18763	11.638%
in0	2120	12.717%	15810	9.806%
citizen. of	1970	11.817%	25742	15.967%
based. in0-x	1882	11.289%	12211	7.574%
citizen. of-x	1844	11.061%	17039	10.569%
member. of	1611	9.663%	19918	12.355%
gpe0	1569	9.412%	18110	11.233%
in0-x	1474	8.842%	8784	5.448%
agent. of	954	5.723%	15776	9.785%
head. of	564	3.383%	7699	4.775%
agency. of	435	2.609%	4775	2.962%
player. of	401	2.405%	5692	3.531%
agency. of-x	382	2.291%	2108	1.308%
head. of state	380	2.279%	7986	4.953%
head. of state-x	343	2.057%	3853	2.390%
appears. in	294	1.764%	4555	2.825%
vs	281	1.686%	7187	4.458%
head. of gov	273	1.638%	4015	2.490%
head. of gov-x	247	1.482%	2383	1.478%
minister. of	234	1.404%	2280	1.414%
minister. of-x	213	1.278%	1629	1.010%
based. in2	185	1.110%	971	0.602%
part. of	164	0.984%	2858	1.773%
in2	157	0.942%	1055	0.654%
created. by	134	0.804%	944	0.586%
agent. of-x	125	0.750%	897	0.556%
award. received	111	0.666%	973	0.604%
institution. of	105	0.630%	2113	1.311%
ministry. of	81	0.486%	666	0.413%
coach. of	65	0.390%	1211	0.751%
won. vs	61	0.366%	1531	0.950%
spouse. of	55	0.330%	599	0.372%
event. in0	51	0.306%	333	0.207%
directed. by	43	0.258%	303	0.188%
spokesperson. of	39	0.234%	177	0.110%
plays. in	38	0.228%	330	0.205%
gpe1	35	0.210%	135	0.084%
product. of	31	0.186%	334	0.207%
event. in2	24	0.144%	175	0.109%
child. of	22	0.132%	281	0.174%
parent. of	22	0.132%	281	0.174%
based. in1	22	0.132%	376	0.233%
signed. by	20	0.120%	521	0.323%
character. in	20	0.120%	47	0.029%
is. meeting	16	0.096%	414	0.257%
law. of	16	0.096%	286	0.177%
in1	13	0.078%	39	0.024%
gpe2	13	0.078%	83	0.051%
advisor. of	10	0.060%	143	0.089%
TOTAL	16671	100.0%	161220	100.0%

Table 13: Relation type statistics.



E RELATION CONSISTENCY RULES

This appendix enumerates the logical predicates used as a consistency check in our dataset.

$$\text{spouse of}(Y, X) \Rightarrow \text{spouse of}(X, Y) \quad (1)$$

$$\text{vs}(Y, X) \Rightarrow \text{vs}(X, Y) \quad (2)$$

$$\text{won-vs}(X, Y) \Rightarrow \text{vs}(X, Y) \quad (3)$$

$$\text{- won vs}(X, Y) \Rightarrow \text{vs}(Y, X)$$

$$\text{-} \quad \text{-} \quad (4)$$

$$\text{child of}(Y, X) \Rightarrow \text{parent of}(X, Y) \quad (5)$$

$$\text{parent of}(Y, X) \Rightarrow \text{child of}(X, Y) \quad (6)$$

$$\text{ministry of}(X, Y) \Rightarrow \text{agency of}(X, Y)$$

$$(7)$$

$$\text{agency of-x}(X, Z) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{agency of}(X, Y)$$

$$\text{-} \quad \text{-} \quad (8) \text{ agency of}(X,$$

$$Y) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{agency of-x}(X, Z) \quad (9)$$

$$\text{agent of-x}(X, Z) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{agent of}(X, Y) \quad (10)$$

$$\text{agent of}(X, Y) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{agent of-x}(X, Z) \quad (11)$$

$$\text{minister of}(X, Y) \Rightarrow \text{agent of}(X, Y) \quad (12)$$

$$\text{- head of gov}(X, Y) \Rightarrow \text{agent of}(X, Y)$$

$$\text{-} \quad \text{-} \quad (13)$$

$$\text{) head of state}(X, Y) \Rightarrow \text{agent of}(X, Y)$$

$$(14$$

$$)$$

$$\text{citizen of-x}(X, Z) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{citizen of}(X, Y)$$

$$\text{-} \quad \text{-} \quad (15)$$

$$\text{citizen of}(X, Y) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{citizen of-x}(X, Z) \quad (16)$$

$$\text{minister of-x}(X, Z) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{minister of}(X, Y) \quad (17)$$

$$\text{minister of}(X, Y) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{minister of-x}(X, Z) \quad (18)$$

$$\text{head of state-x}(X, Z) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{head of state}(X, Y)$$

$$\text{-} \quad \text{-} \quad (19)$$

$$\text{head of state}(X, Y) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{head of state-x}(X, Z)$$

$$(20)$$

$$\text{head of gov-x}(X, Z) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{head of gov}(X, Y)$$

$$(21) \text{ head of gov}(X, Y$$

$$) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{head of gov-x}(X, Z) \quad (22)$$

$$\text{in0-x}(X, Z) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{in0}(X, Y) \quad (23)$$

$$\text{in0}(X, Y) \wedge \text{gpe0}(Z, Y) \Rightarrow \text{in0-x}(X, Z) \quad (24)$$

$$\text{in2}(X, Z) \wedge \text{in0}(Z, Y) \Rightarrow \text{in0}(X, Y) \quad (25)$$

$$\text{in1}(X, Z) \wedge \text{in0}(Z, Y) \Rightarrow \text{in0}(X, Y) \quad (26)$$

$$\text{based in2}(X, Z) \wedge \text{in0}(Z, Y) \Rightarrow \text{based in0}(X, Y)$$

$$\text{-} \quad \text{-} \quad (27)$$

$$\text{based in1}(X, Z) \wedge \text{in0}(Z, Y) \Rightarrow \text{based in0}(X, Y) \quad (28)$$

$$\text{agency-of}(X, Y) \wedge \text{type} :: \text{gpe0}(Y) \Rightarrow \text{based in0}(X, Y)$$



(29)

$$\text{event in2}(X, Z) \wedge \text{in0}(Z, Y) \Rightarrow \text{event in0}(X, Y) \quad (30)$$

$$\text{event in1}(X, Z) \wedge \text{in0}(Z, Y) \Rightarrow \text{event in0}(X, Y)$$

(31)

$$\text{head-of}(X, Y) \Rightarrow \text{member of}(X, Y) \quad (32)$$

$$\text{coach of}(X, Y) \Rightarrow \text{member of}(X, Y)$$

(33)



- spokesperson of(X, Y) \Rightarrow member of(X, Y) (34)
- member of(X, Y) \wedge sport player(X) \Rightarrow player of(X, Y) (35)
- mayor of(X, Y) \Rightarrow head of gov(X, Y) (36)
- directed by(X, Y) \Rightarrow created by(X, Y) (37)
- character in(X, Y) \wedge played by(X, Z) \Rightarrow plays in(Z, Y) (38)
- institution of(X, Y) \Rightarrow part of(X, Y) (39)
- based in0-x(X, Z) \wedge gpe0(Z, Y) \Rightarrow based in0(X, Y) (40)
- based in0(X, Y) \wedge gpe0(Z, Y) \Rightarrow based in0-x(X, Z) (41)

Annex 2: NEM accepted abstract



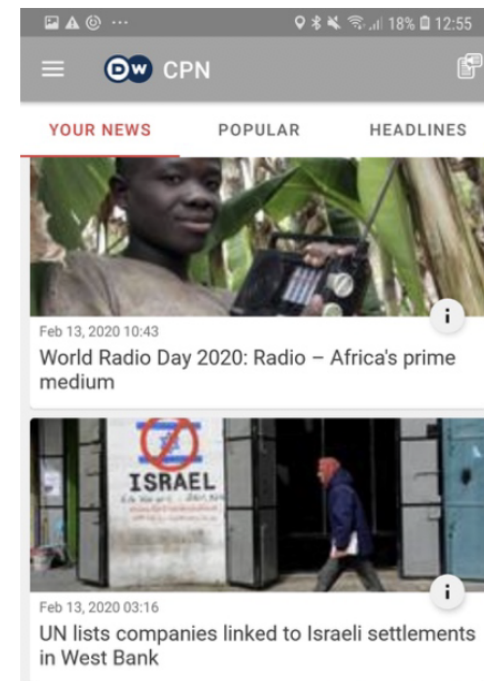
'Have I got news for you!' The challenges of news personalisation from an end user perspective

Today many media users are confronted with an overflow of audiovisual content and information on different platforms, making them feel overwhelmed. For end users it has become difficult to determine which news is relevant and trustworthy, as a lot of news circulating online can be categorized as 'fake news' and 'misinformation'. Not every news consumer has the same level of knowledge on or interest in a topic, leading to a need for diversification. Furthermore, there is the danger of the so-called "filter bubble" which might lead to a biased and unnuanced view on important societal topics (Pariser, 2011)¹. Traditional news organizations have been a trusted news source for media users over the past decades. In today's society where information has increased exponentially, there is an important role for these established media companies to offer relevant and trustworthy news to their audiences².

To address these challenges, the European Horizon 2020 project CPN (Content Personalisation Network) aims to offer media consumers more relevant news, in the right format and at the right time. To achieve this aim, a personalized news application was developed with content from its media partners VRT (Belgium), Dias (Cyprus) and Deutsche Welle (Germany). The application offers personalized news in three languages: English (DW), Greek (Dias) and Dutch (VRT). This news application has been developed together with all relevant stakeholders including the users in a user-centered design approach. The application has been iteratively developed and tested in three pilot rounds of which the last round is an open pilot with hundreds of participants taking place February-March 2020.

In this contribution we will discuss the end user evaluation of the CPN app, focusing on specific elements such as the user acceptance towards news personalisation, the issue of fear of missing out (FOMO), the sustainability of the notion of filter bubble from an end user perspective and whether there is an increased level of the feeling of being informed after using the CPN application. The discussed results are based on a combination of quantitative data (logging data and survey data) and qualitative data (interviews with test-users).

¹ Pariser, Eli (2011). *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Press. ISBN 978-1-59420-300-8



CPN app (third prototype)

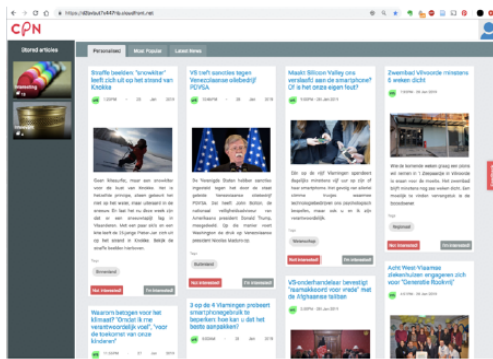
The first two pilot rounds already provided relevant insights related to news personalisation. In 2018, almost 100 end-users tested the first prototype of the CPN news recommender platform for a period of 10 days. Specifically, the participating end-users tested the web interface of the recommender. In each pilot country, the recommender system contained news content from the local media partner (VRT, Deutsche Welle and DIAS).

The [first CPN prototype](#) consisted of three news sections, presented in separate tabs:

- **Most popular** - the most-read content among the users of the particular news outlet
- **Latest news** - all articles in chronological order, with the latest on top
- **Personalised** - the most relevant content for the user

² <https://www.reuters.com/article/us-media-fakenews/fake-news-hurts-trust-in-media-mainstream-outlets-fare-better-poll-idUSKBN1D002S>





First CPN prototype

The ‘personalised’ tab was the most read section in all three pilot countries. While news personalisation itself was evaluated as positive, there was a big fear of missing out (FOMO) when receiving personalised news articles: the participants were afraid that through personalisation they might only get news content based on their interests, which could cause them to miss other relevant or important news. The CPN recommender was not available as a mobile application during the first pilot, which was a negative point among the testers. On the other hand, the participants generally thought the web interface was straightforward and easy to use, although some thought that more important articles could be made more prominent in the interface.

We tested the [second prototype](#) which was an app in 2019. We evaluated whether users liked the CPN personalisation algorithm by asking them how informed they felt after using our prototype. VRT, DIAS and Deutsche Welle, the media partners in the project, invited users to download the (Android) app onto their mobile phones and follow the news through CPN for a test period of four weeks.

News articles in the new mobile app could be found under three different tabs:

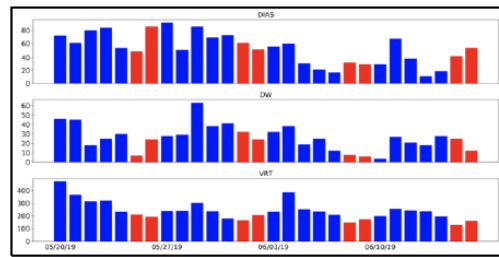
- **My news:** personalised articles
- **Headlines:** articles selected by the news department
- **Just in:** most recently published articles

Authors

Wendy Van den Broeck (wendy.van.den.broeck@vub.be), imec-SMIT, Vrije Universiteit Brussel; Natasja Van Buggenhout, imec-SMIT, Vrije Universiteit Brussel (natasja.van.buggenhout@vub.be); Ilke Lemmelijn, VRT (ilke.lemmelijn@vrt.be); Joris Mattheijssens, VRT (joris.mattheijssens@vrt.be); Jens Van Lier, VRT (jens.van.lier@vrt.be).

What did we learn?

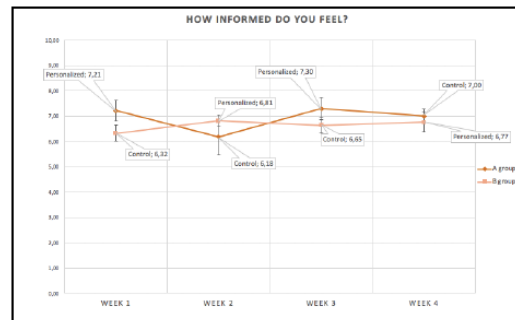
We monitored and compared how users experienced both a personalised and non-personalised offer throughout the whole test. The users were split into two halves: one group started out with just a random selection of articles, while the other received real personalized news recommendations. We switched the groups weekly and checked whether users could tell the difference between the two versions and which one they liked better.



How often did users click on articles during the evaluation?

People did appreciate the recommended results over the random results, felt more informed and gave us positive feedback on the application - but overall, statistically the differences were too slim to clearly say the CPN app fully convinced testers. That’s why in the third and final pilot, the user-base for testing will be increased up to a minimum of 500 users in total.

The project will finish end of April so all project results will be available for presentation during NEM 2020.



How informed did users feel? More on the personalised side.

