PHDD: A Corpus of Physical Health Data Disclosure on Twitter During the COVID-19 Pandemic

DPVCG Meeting, Wednesday, February 17, 2021

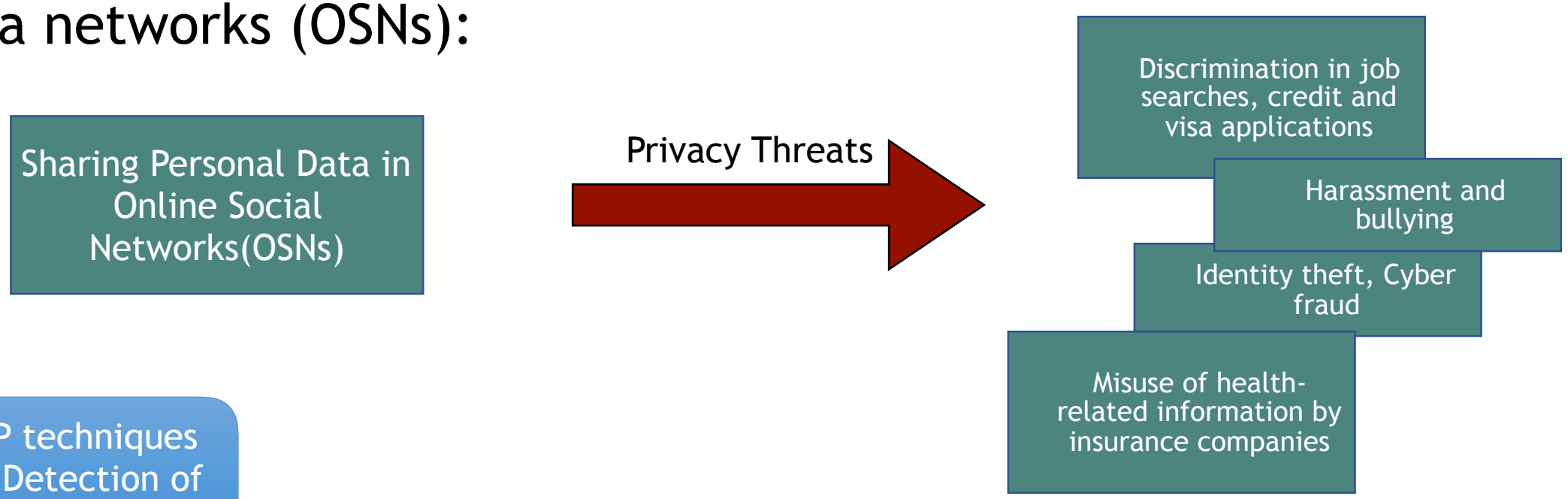
Rana Saniei
Víctor Rodríguez-Doncel

Ontology Engineering Group – Universidad Politécnica de Madrid

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°813497

Introduction

- Nowadays many users share their personal data in online social media networks (OSNs):

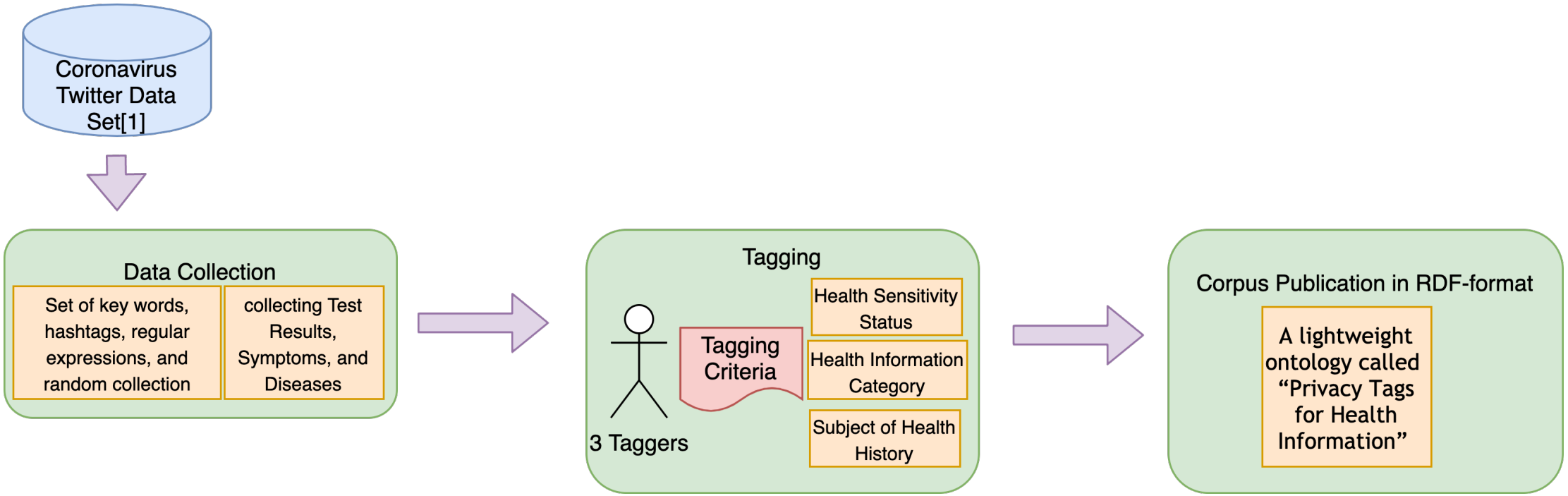


NLP techniques for Detection of Personal Data

detecting personal data or PIIs in textual documents is one of the important challenges that have not been fully overcome yet!

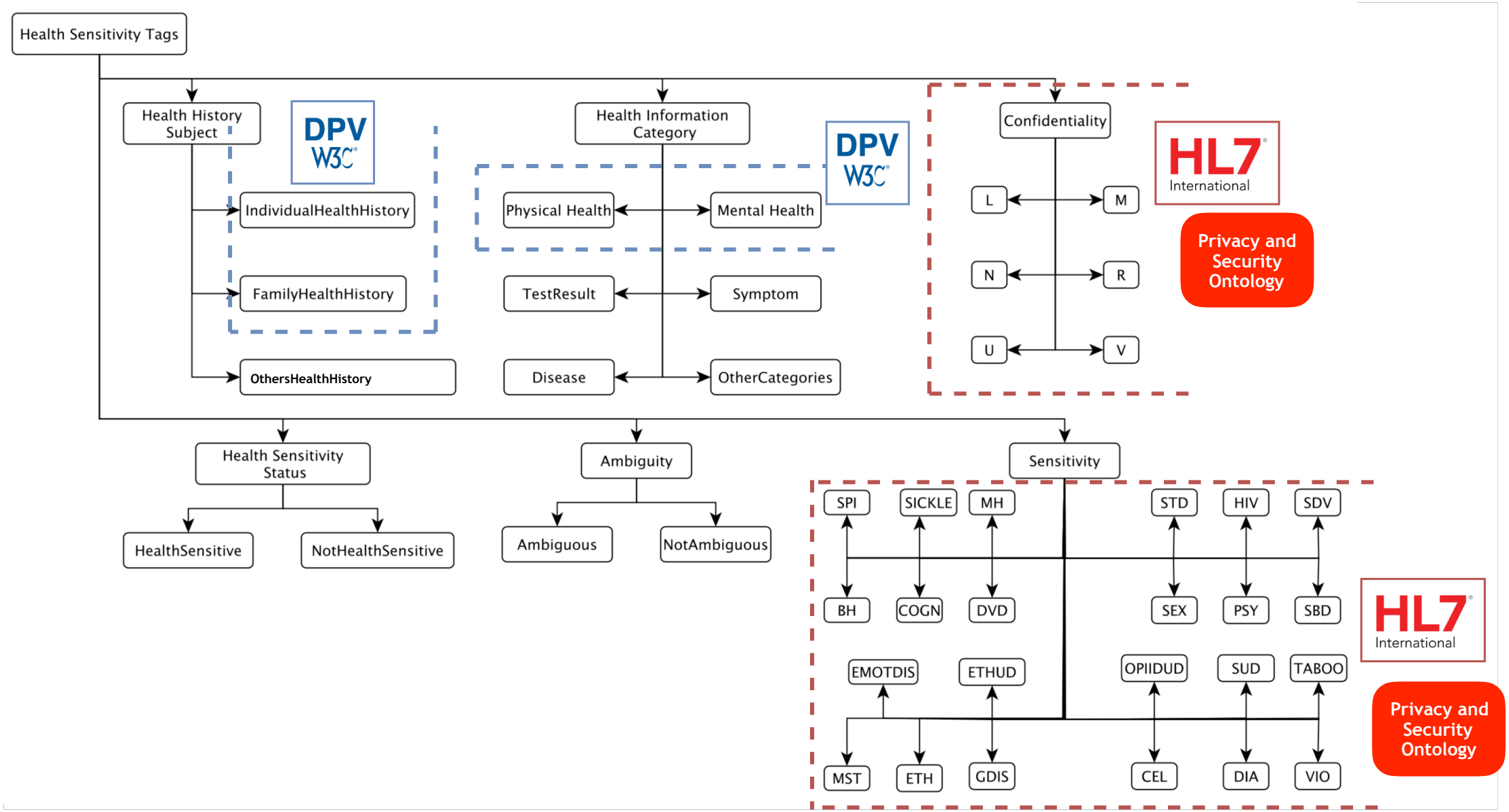
- Main objective: Building a Corpus of **Physical Health Data Disclosure on Twitter During COVID-19**

PHDD: A Corpus of **P**hysical **H**ealth **D**ata **D**isclosure on Twitter During the Covid-19 Pandemic



[1]Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set

Ontology of Privacy Tags for Health Information

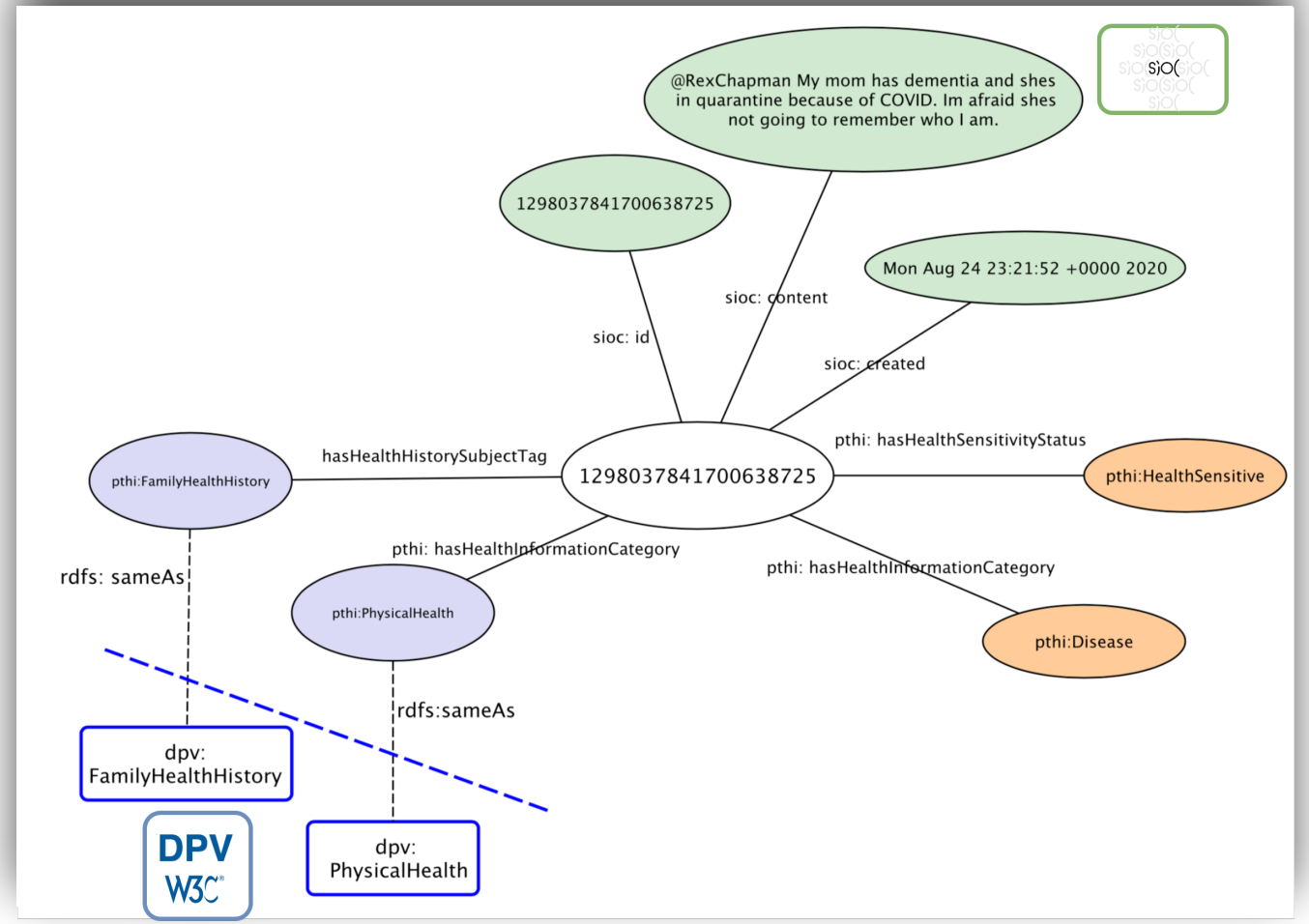


Final corpus statistics

	Total	Average
Tweets	1,494	-
Sentences	5,126	3.42
Tokens	63,029	42.15
Hashtags	685	0.45
Mentions	855	0.57
URLs	537	0.36
Verbs	7,437	4.97
Nouns	11,114	7.43
Proper Nouns	3,134	2.10
Pronouns	6,778	4.53
Adjective	3,867	2.58
Adverbs	3,143	2.10

Health Sensitivity Status		
HS	656	41.12%
NHS	797	47.86%
Ambiguous	38	1.13%
Health Information Category		
TES	202	30.03%
SYM	192	47.86%
DIS	346	51.30%
OTH	51	6.00%
Health History Subject		
IND	351	56.49%
FAM	244	38.96%
OTH	52	7.14%

A Sample Record in the Corpus



- Tagging criteria and final corpus are publicly available on https://protect.oeg.fi.upm.es/def/phdd/.

Future works

- Implementation of a **supervised ML technique** to detect health-sensitive information in textual inputs.
  - Notify users if their shared content contains any health-sensitive information.
  - Implementation of a **fine-grained access control mechanism** which lets people to define fine-grained privacy preferences:

"I do not want the medical information about my family members, revealed by me on Twitter, being shared to any other data processors/ joint controllers"

"I want Twitter to delete permanently all of my health-related information from their databases."

Protect

QUESTIONS

Rana Saniei
Email: r.saniei@upm.es
Twitter: @RanaSaniei
Víctor Rodríguez-Doncel

Thanks to Delaram Golpaygani, Beatriz Esteves, and Karen Vázquez Flores for their collaboration in annotating the corpus.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°813497