Data Privacy in Journalistic Knowledge Platforms

Marc Gallofré Ocaña[a], Tareq Al-Moslmi[a] and Andreas L. Opdahl[a]

[a]*University of Bergen, Fosswinckelsgt. 6, Postboks 7802, 5020 Bergen, Norway*

Abstract

Journalistic knowledge platforms (JKPs) leverage data from the news, social media and other sources. They collect large amounts of data and attempt to extract potentially news-relevant information for news production. At the same time, by harvesting and recombining big data, they can challenge data privacy ethically and legally. Knowledge graphs offer new possibilities for representing information in JKPs, but their power also amplifies long-standing privacy concerns. This paper studies the implications of data privacy policies for JKPs. To do so, we have reviewed the GDPR and identified different areas where it potentially conflicts with JKPs.

Keywords

Privacy, Personal data, Journalistic Knowledge Platforms, GDPR

1. Introduction

Journalistic Knowledge Platforms (JKPs) are an emerging generation of platforms which combine state-of-the-art artificial intelligence (AI) techniques, like knowledge graphs and natural-language processing (NLP) [1, 2] for transforming newsrooms and leveraging information technologies to increase the quality and lower the cost of news production. JKPs exploit and combine news, social media and other information sources, using linked open data (LOD), digital encyclopaedic sources and news archives to construct knowledge graphs and provide fresh and unexpected information to journalists, helping them dive more deeply into information, events and story-lines [3]. JKPs of various kinds are becoming increasingly important in leading news agencies like BBC [4] and Thomson Reuters [5].

However, obtaining and representing knowledge leads to data privacy concerns when personal data from different sources is neither collected directly from the subject nor with the subject's consent, although some countries have exemptions that loosen privacy requirements for journalistic research that is in the public interest or does not identify individuals directly. This exemption becomes even more complex when the national privacy policies that apply to the data sources and the JKP are distinct or the public

interest is not crystal clear.

Data privacy has become a central topic of discussion for organisations and projects from private companies and governments to research activities in universities around the globe. Whereas there is no general solution to privacy for everyone and specific solutions vary between different countries, cultures and organisations, privacy is a common concern, which has been discussed from the ethical and philosophical points of view by many different authors [6, 7] and organisations like the European Commission [8, 9]. The EU has established the General Data Protection Regulation (GDPR) which sets up a framework for governing the usage, processing, privacy and security of personal data, granting individuals power over their data and making organisations responsible for data collection and usage practices.

Our group have been developing News Hunter [10, 11, 12], a series of JKP architectures and prototypes. The current News Hunter platform is big-data ready and designed to continually harvest and monitor real-time news feeds (e.g., RSS or web-sites) and social media (e.g., Twitter and Facebook). It aims to analyse and represent news content semantically in knowledge graphs in order to provide better background information for journalists and to suggest news angles [13, 14, 15, 16, 17].

As part of our News Hunter effort, this paper investigates the implications of the GDPR on JKPs. To do so, we asked ourselves which data privacy conflicts can arise when JKPs when are used in journalistic work, in particular when that work may be exempted from some privacy regulations because it is in the public interest. To the best of our knowledge, there is no previous work discussing the possible data privacy conflicts in JKPs. Our contributions are: (1) we review different journalistic scenarios and personal data sources that

can conflict with GDPR policies, and (2) we introduce a personal data matrix framework to classify personal data conflicts and discuss the possible uses of this matrix.

This paper is organised as follows: section 2 defines the main privacy concepts, section 3 discusses potential data privacy conflicts in JKPs, section 4 introduces the personal data matrix framework, section 5 summarises the conclusions, and section 6 presents open questions and future work.

2. Background

2.1. Journalistic knowledge platforms

Journalistic Knowledge Platforms (JKPs) leverage and combine news, multimedia content (e.g., TV news channels and podcast) social media (e.g., Twitter and Facebook), web-blogs and information over the net, using linked open data (LOD), digital encyclopaedic sources (e.g., Wikipedia and Wikidata) and news archives to provide fresh and unexpected information to journalists. Projects like Neptuno [18], Event Registry [19], NEWS [20], NewsReader [21], SUMMA [22] and News Angler [11, 16, 12] have presented examples of JKPs.

A typical JKP comprises a knowledge graph [23, 24, 25] along with AI, NLP pipelines, and semantic technology components. In a JKP, the Knowledge graph is filled with potentially news-related histories, information and current and archival news to support journalists in creating newsworthy stories, finding relevant information, events and story-lines, and validating and verifying news. The information in the knowledge graph is represented using standard identifiers and semantic knowledge representations with reasoning capabilities. The usage of standard identifiers facilitates data integration, which is the process of joining and merging different data sets or public data sources like Semantic Web [26, 27], linked open data (LOD) [28], including Wikidata and DBpedia, and Wikipedia. Data integration together with reasoning allows drawing new insights from information from across the data that would be impossible before with isolated datasets. This inherent ability of drawing new insights implies that new personal data may be derived and exposed in the knowledge graph.

2.2. Privacy

Privacy is a historically and culturally situated concept. For example, whereas privacy in Europe is traditionally considered as an inalienable basic right of an autonomous person that states must protect to preserve a democratic society, the concept of privacy in the United States of America is understood primarily as a physical notion that implies the "private space" (e.g., bedroom, bathroom or the entire home) [6]. These differences are reflected in the data privacy regulations of the EU and the USA. The EU states in the *General Data Protection Regulation* (GDPR) [29] that individuals must be notified and have the right to consent when their personal data is collected from either inside or outside EU legislation. In contrast, the US only regulates privacy issues regarding health matters and some financial information, leaving the rest to individual states or businesses which do not need to ask for individuals consent and give the possibility to individuals to resign if they have any reservations about what is being collected from them. On a global scale beyond EU and the USA, differences in how privacy is viewed are even bigger, making it even more challenging to handle privacy regulations when JKPs are used in fully international news organisations that operate across cultural and legal domains.

2.3. GDPR

All actions using or processing personal data of data subjects who are in the European Union shall obey the General Data Protection Regulation (GDPR) [29]. The GDPR is an extensive regulation which sets the basis for dealing with personal data in the EU or using personal data from the EU. This section highlights the most general concepts that restrict what and how to process personal data in JKPs.

The GDPR defines the concepts of *personal data* and *processing* (*Chapter I, Article 4*). *Personal data* is any information that can be employed to identify directly or indirectly a natural person (e.g., name, an identification number or online identifier) or *sensitive data* like health, biometric, genetic, economic, cultural factors or political opinions of a natural person. Data that has been de-identified, encrypted or pseudonymised but can be used to re-identify a person is considered as personal data too. By *processing*, the GDPR means processes such as collection, structuring, storage, alteration, consultation, use, disclosure, combination, restriction, erasure or destruction of personal data.

Moreover, the GDPR establishes a set of principles for processing personal data (*Chapter II*) which define how data have to be processed, stored and maintained. These set of principles establish that data shall be processed within the initial purposes and purposes compatible with them (*purpose limitation*), only what is necessary to the purpose (*data minimisation*), personal

data shall be accurate and kept up to date (*accuracy*), and stored for no longer periods than the necessary for the purpose (*storage limitation*). It also defines the *lawfulness of processing* which determines when personal data can be processed, e.g., when data subject gives the consent or for a task carried out in public interest. Under the GDPR, some research by journalist and academics is understood as public interest. Likewise, the GDPR limits the processing of sensitive data which is prohibited in general terms but with some exceptions, e.g., when data subject gives explicit consent, it is necessary for reasons of substantial public interest, or the data subject has manifestly made it public.

The GDPR also details when and which information have to be provided to the data subjects (*Chapter III*). In the case of personal data that is not obtained directly from the subject, it determines which data have to be provided, e.g., the source of the personal data and whether it came from publicly accessible sources or the categories of personal data. Nevertheless, it also establishes some exemptions, e.g., when the provision of such information proves to be impossible or is likely to harm the objectives of the processing objective.

3. Privacy conflicts in JKPs

When discussing which scenarios in JKPs can cause a conflict with the GDPR we must consider the source of the personal data, distinguishing between the data gathered directly from the subject, the data harvested from other sources like news or social media and the inferred data.

In the context of GDPR, some data processing by journalists is exempted when it is conducted in the public interest. However, this exemption exclusively applies to journalistically relevant (newsworthy) personal data, not to any personal data processed in the JKP, and sensitive data may be less exempted or not exempted at all. Therefore, we must also consider how relevant the personal data is for the public interest from a news perspective. This includes the assessment of newsworthiness [30] along with the type of news. E.g., a corruption scandal and a private event in the life of a famous person may both be highly newsworthy, but corruption is most likely more important for the public interest.

3.1. Personal data from the subject

When personal data comes directly from a subject and is collected with the subject's explicit consent (e.g., personal data collected during an interviewed), it does not present a problem with the GDPR. However, the data can be made it publicly accessible by the subject itself in social media networks (e.g., posts like tweets in Twitter or forums and groups like Facebook groups) or in the subject's verifiable social media accounts and personal web sites without providing explicit consent for its collection to the JKP. In that case, apart from having to follow the source's data policies, it raises the ethical questions whether the consent is implicit because it is publicly available, when we should consider that it is publicly available and under which conditions.

3.2. Personal data from third parties

When the data is not collected directly from the subject, instead, it has been made accessible by a third party and subject may ignore its existence, we have to consider two possible scenarios:

The first scenario, when news-related information is gathered from the web (e.g., online news, RSS, websites or social media), JKPs can extract personal data from the content to represent and combine it in the knowledge graphs. E.g., from *"We know the classic 7-layer dip, made with Bush's Beans, is a fan favorite for game day snacking celebrations, Kate Rafferty, the consumer experience manager for Bush's, told Fox News."*[1], we can extract information like "Kate Rafferty is a person who works as consumer experience manager at Bush's Beens company at Knoxville, Tennessee" which can be considered as personal data, as it can be used to identify a natural person. According to the GDPR, the subject should be notified, and the JKP has to provide a mechanism for the subject to protest. Even though, on a large scale, two issues arise: the number of notifications that famous people will get and how to contact subjects if the content information is missing.

The second scenario, when personal data is gathered from publicly available sources or open sources like Wikipedia, Wikidata or telephone/address books, it is clear that the personal data is already public. However, it may not be released with subject's consent. In that case, it opens the question about: why should the JKP not be allowed to store copies of personal data which is already public?

3.3. Inferred personal data

When personal data is not gathered from any source, instead, it is inferred using the actual data (either from the data subject, collected from news or gathered from

---

[1]Drew Schwartz, VICE: This 70-Layer Bean Dip Is the Most Vile Thing I've Ever Seen (https://t.co/qKyyNevpBh)

public sources) and reasoning techniques. E.g., from the text *"The European Court of Justice (ECJ) said that Oriol Junqueras had become an MEP the moment he was elected in May, despite being on trial for sedition."*[2], we can represent the person "Oriol Junqueras" as the entity Q116812 from Wikidata, from which we can derive that he is a member of a political party (P102) and the political party is "the Republican Left of Catalonia" (Q150068). With this information is it possible to infer the subject's political ideology (P1142) from the political party information such as "republicanism" (Q877848) and "Catalan pro-independence movement" (Q893331). In this scenario there is not a direct source of subject's personal data or political opinions, instead, there is a source of related information used for inferring knowledge which can be either in the same knowledge graph or from external sources.

3.4. Possible solutions

To comply with the GDPR's *Chapter III*, in any of the previously discussed scenarios it is important to identify the data source and personal data category (e.g., name, ID number, online identifier, health data, political opinion). Thus, it will be possible to identify both the source and data and take actions accordingly. Although the main responsible of complying with the GDRP in the first place is the data provider (i.e., news website, social media platform or telephone/address books), JKPs should follow the GDPR to safeguard the subjects of privacy and consider the policies and restrictions established by the data provider. The JKP must always take independent responsibility for privacy, and it cannot trust its sources to safeguard privacy. In a truly international and global set-up, where different privacy policies apply, JKPs may have to be designed with different knowledge graphs for different legal domains or geographical regions, each graph only being accessible from its own privacy domain. When this is infeasible, the most restrictive policies to guarantee personal data privacy must be adopted.

Moreover, JKPs should also implement automatic mechanisms to notify subjects with both the personal data and the sources when this information is identified, a process that can be done by email. It is also possible to set up an automatic system for subjects to protest, complain, request or ask about personal data.

Table 1
Personal data matrix

	Consented	Collected	Inferred
Impersonal Data	✔	✔	✔
Personal Data	✔	!	!
Sensitive Data	✔	‼	‼

4. Personal data matrix

After reviewing the previous scenarios, we classified the different situations that can cause a conflict with the GDPR into a two-dimension matrix (figure 1) framework. The personal data matrix aims to help journalists and JKP developers to classify the personal data in JKPs and its possible issues with privacy policies.

The personal data matrix (figure 1) classifies personal data based on the privacy level and the data source. The first dimension (privacy level) classifies the data whenever it does not represent personal data (impersonal data), it represents personal data or it represents "sensitive data". There is an explicit distinction between personal and "sensitive data" because in the GDPR "sensitive data" have much more restricted limitations. The other dimension, the data source dimension, classifies data based on the data collected with the subject's consent (consented), the data directly collected from the content and the inferred data. Only when data is either explicitly consented or it is not personal data its treatment is straightforward. Otherwise, as discussed in the previous section, each of these combinations has its issues and open questions regarding the application of the GDPR and its origin.

The data matrix can be also regarded as a cube, where the public interest represents the third dimension. This third dimension determines to what extent the GDRP exemption to data processing for journalistic purposes in the public interest applies, taking into consideration the newsworthiness component of the data.

The proposed matrix can be used by JKP researchers and developers to ensure – as automatically as possible, but in practice aided by human data privacy stewards – that privacy regulations are never violated. The matrix should be used in the design of JKPs to ensure that personal data is protected by default. E.g., developers of JKPs can use the matrix to evaluate the system and identify which processes or collected data can lead to privacy conflicts; implement the matrix as part of the news creation workflow so that journalists can automatically check data privacy compliance before collecting, re-combining or using any personal data;

---

[2]BBC: Jailed Catalan leader 'should have had immunity', rules EU court (https://www.bbc.com/news/world-europe-50808766)

it can be utilized as metadata for each piece of data in the knowledge graph to automatise its recognition and privacy assurance; and the matrix can be used when dealing with data under different regulations to find divergences between them.

5. Conclusion

JKPs need to deal with personal data which in many cases will be integrated into knowledge graphs without the explicit consent from the subject. Thus, JKPs need to safeguard data privacy. For that reason, we have presented a framework for classifying personal data in journalistic knowledge graphs and identified different scenarios and personal data sources that potentially can conflict with the GDPR. We believe the identified scenarios, sources and presented matrix will be helpful as a reference for related projects and similar domains.

6. Future work

We want to continue exploring the open questions highlighted in our discussions in section 3, as well as questions such as how to deal with different privacy regulations that may apply in international settings, how to represent and effectively use GDPR in JKP processes, and how to deal with personal data about children. Data linking transparency is another open question which would help to identify situations that conflict with privacy and identify which data can be stored and which data cannot be stored in JKPs according to the GDPR and other privacy policies and regulations. Besides that, as anonymisation, encryption and blockchain technologies are presented as potential solutions to safeguard privacy and control copyrights and data access, we want to research how effective they are in the context of JKPs and how they can benefit JKPs.

Apart from that, one critical aspect when dealing with data from external sources, which has not been considered in this work, is the copyright and intellectual property regulations which have a direct relation with the data that can be processed and stored. In this context, we want to explore how to effectively manage them in JKPs (e.g., using ontologies).

Acknowledgments

References

[1] T. Al-Moslmi, M. Gallofré Ocaña, Lifting news into a journalistic knowledge platform, in: Proceedings of the CIKM 2020 Workshops, Galway, Ireland, 2020. To appear.

[2] T. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl, C. Veres, Named entity extraction for knowledge graphs: A literature overview, IEEE Access 8 (2020) 32862–32881.

[3] M. Gallofré Ocaña, A. L. Opdahl, Challenges and opportunities for journalistic knowledge platforms, in: Proceedings of the CIKM 2020 Workshops, Galway, Ireland, 2020. To appear.

[4] Y. Raimond, M. Smethurst, A. McParland, C. Lowis, Using the past to explain the present: Interlinking current affairs with archives via the semantic web, in: H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, K. Janowicz (Eds.), The Semantic Web – ISWC 2013, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 146–161. doi:10.1007/978-3-642-41338-4_10.

[5] B. Ulicny, Constructing knowledge graphs with trust, in: METHOD 2015: The 4th International Workshop on Methods for Establishing Trust of (Open) Data, Bethlehem, PA, 2015.

[6] C. Ess, Digital Media Ethics, Polity, 2014.

[7] D. G. Johnson, Computer Ethics, Prentice Hall, 2001.

[8] European Parliament, Regulation (EU) no 1291/2013 of the european parliament and of the council of 11 december 2013 establishing horizon 2020 - the framework programme for research and innovation (2014-2020) and repealing decision no 1982/2006/ECText with EEA relevance (2013).

[9] European Commission, Policy | science with and for society - research and innovation - european commission, 2019. URL: http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=ethics.

[10] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, News hunter: building and mining knowledge graphs for newsroom systems, in: NOKOBIT—Norsk konferanse for

organisasjoners bruk av informasjonsteknologi, volume 26, 2018.

[11] M. Gallofré Ocaña, L. Nyre, A. L. Opdahl, B. Tessem, C. Trattner, C. Veres, Towards a big data platform for news angles, in: 4th Norwegian Big Data Symposium (NOBIDS) 2018, 2018.

[12] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, A knowledge graph platform for newsrooms, Computers in Industry (2020). To appear.

[13] B. Tessem, A. L. Opdahl, Supporting journalistic news angles with models and analogies, in: 2019 13th International Conference on Research Challenges in Information Science (RCIS), IEEE, 2019, pp. 1–7. doi:10.1109/RCIS.2019.8877058.

[14] A. L. Opdahl, B. Tessem, Towards ontological support for journalistic angles, in: Enterprise, Business-Process and Information Systems Modeling, Springer International Publishing, 2019, pp. 279–294. doi:10.1007/978-3-030-20618-5_19.

[15] B. Tessem, Analogical news angles from text similarity, in: Artificial Intelligence XXXVI, Springer International Publishing, 2019, pp. 449–455. doi:10.1007/978-3-030-34885-4_35.

[16] A. L. Opdahl, B. Tessem, Ontologies for finding journalistic angles, Software and Systems Modeling (2020) 1–17.

[17] E. Motta, E. Daga, A. L. Opdahl, B. Tessem, Analysis and design of computational news angles, IEEE Access (2020).

[18] P. Castells, F. Perdrix, E. Pulido, R. Mariano, R. Benjamins, J. Contreras, J. Lorés, Neptuno: Semantic web technologies for a digital newspaper archive, in: European Semantic Web Symposium, Springer, Berlin, Heidelberg, 2004, pp. 445–458.

[19] G. Leban, B. Fortuna, J. Brank, M. Grobelnik, Event registry: Learning about world events from news, in: Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, ACM, New York, NY, USA, 2014, pp. 107–110. doi:10.1145/2567948.2577024.

[20] N. Fernández, J. M. Blázquez, J. A. Fisteus, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, Z. Ben-Asher, News: Bringing semantic web technologies into news agencies, in: I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. M. Aroyo (Eds.), The Semantic Web - ISWC 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 778–791.

[21] P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. Palmero Aprosio, G. Rigau, M. Rospocher, R. Segers, NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news, Knowledge-Based Systems 110 (2016). doi:10.1016/j.knosys.2016.07.013.

[22] U. Germann, P. v. d. Kreeft, G. Barzdins, A. Birch, The summa platform: Scalable understanding of multilingual media, in: Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 2018.

[23] A. Singhal, Introducing the knowledge graph: things, not strings, 2012. URL: https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

[24] Ehrlinger, Lisa and Wöß, Wolfram, Towards a definition of knowledge graphs, in: Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), 2016, p. 4. URL: http://ceur-ws.org/Vol-1695/paper4.pdf.

[25] J. Yan, C. Wang, W. Cheng, M. Gao, A. Zhou, A retrospective of knowledge graphs, Frontiers of Computer Science (2016). doi:10.1007/s11704-016-5228-9.

[26] T. Berners-Lee, J. Hendler, O. Lassila, et al., The semantic web, Scientific American 284 (2001).

[27] N. Shadbolt, T. Berners-Lee, W. Hall, The semantic web revisited, IEEE Intell. Syst. 21 (2006) 96–101.

[28] C. Bizer, T. Heath, T. Berners-Lee, Linked data: The story so far, in: Semantic services, interoperability and web applications: emerging concepts, IGI Global, 2011, pp. 205–227.

[29] The European Parliament and The Council of the European Union, Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), Official Journal of the European Union (2016). URL: http://data.europa.eu/eli/reg/2016/679/oj.

[30] T. A. A. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl, B. Tessem, Detecting newsworthy events in a journalistic platform, in: The 3rd European Data and Computational Journalism Conference, 2019, pp. 3–5.