**Research data management plan (RDMP)**

|  |  |  |
| --- | --- | --- |
| Administrative information | | |
|  | Principal investigator | professor Nebojša Blanuša | |
|  | Affiliation | University of Zagreb – Faculty of political science | |
|  | Project proposal title | **Researching Europe, Digitalisation, and Conspiracy Theories - REDACT** | |
|  | RDMP contact person | Nebojša Blanuša (nebojsa.blanusa@fpzg.hr) | |
| 1. | Data collection and documentation | | |
|  | What data will you collect, analyse, generate or reuse? (Please state the type, format and volume of data you will collect, not only final data set that will be the result research) | This project will create four main types of data: (1) a selection of thematically significant text, image and audio-visual materials collected from websites and social media platforms, to be used as the basis for the team’s discourse analysis (stored as RTF and PDF files for text and image, with links to audio-visual materials as necessary, c. 200GB); (2) audio recordings and transcripts of interviews with organisations tasked with combatting disinformation (c. 50GB) (audio stored in MP3 format and transcribed using Nvivo software, c. 100GB); (3) big data searchable archives of individual social media platforms, collected using custom scraping tools (see below, c. 10TB); and (4) data visualisations resulting from those big data queries (created using Gephi, c. 50GB). | |
|  | How will the data be collected, processed, or generated? (Briefly describe methodologies and quality assurance processes you will use, organization of your project files and data, tools and instruments which will be used for collecting and processing the data) | The data will be collected from digital social media and conducting ethnographic work within factchecking organisations. REDACT will utilise both the macro approach of data analytics with the micro attention of cultural studies, digital ethnography (Kozinets 2019) and ethnography. This methodological combination amounts to ‘data hermeneutics’ (Gerbaudo 2016), an approach that Birchall and Knight first trialled to good effect in a 1- year UKRI-funded project on online conspiracy theories concerning the coronavirus pandemic. In the first stage of this data hermeneutics approach, the UK-based Postdoc will draw on the local expertise and subject knowledge of the REDACT team to construct region-specific seed lists of search terms for subsequent digital queries, to identify relevant social media groups and content producers as well as other websites and blogs in a way that is sensitive to the specific cultural, political and media contexts of each European region (objective 1).  These methods will be used to identify the most engaged-with conspiracy theory content and most active communities in our case study countries, from both mainstream moderated and unmoderated social media platforms as well as other online forums. The methods will include API capture and analysis toolkits such as CrowdTangle and 4CAT; NLP, topic modelling, named-entity recognition and the narrative analysis techniques that Tangherlini (2020) and his team have successfully used in relation to online conspiracism; and social network analysis and other data visualisation tools such as Gephi. While digital ‘distant reading’ is commonplace in the social sciences, it often side-lines cultural specificity and questions of identity when analysing online discourse and sociality.  At the same time, those forms of hermeneutic close reading practiced in the humanities are incapable of addressing the huge volume and variety of online activity. To avoid these shortcomings, the goal of data hermeneutics is to create manageable and meaningful datasets that allow researchers to conduct qualitative research, or ‘close data reading’ (Gerbaudo 2016), on carefully curated samples of social media posts and other forms of online conspiracism. This approach will allow us to situate social media posts within both the conversations taking place on particular platforms and the wider discourses, narratives and worldviews that shape conspiracy theories. A range of techniques will be employed to produce manageable and meaningful datasets, including top sampling (focusing on the most engaged with hashtags, keywords, posts and/or posters for each platform, channel or website); random sampling (assembling representative examples of particular conversations, topics or memes from the larger dataset); and zoom-in sampling (concentrating on a particularly significant conversation or theme). These datasets will then be manually cleaned and coded by the PIs and RAs on the project, using constant comparison to remove false positives and identify significant recurring themes. Once these curated datasets have been created and cleaned, the regional teams will conduct digital ethnographies of the sociality around these posts and digital spaces, paying particular attention to how issues of gender, ethnicity and class feature and are negotiated in conspiracist online exchanges.  As a general rule, we are interested in what people say and do online and not in who they are, and will not  follow them to their individual Facebook pages or other profiles.  Given this focus on what conspiracysympathetic users do online—on what digitalisation affords them—it is more fitting to conduct observational digital ethnography rather than in-person ethnography. We will combine this digital ethnography with critical discourse analysis (Fairclough 2010) and a semiotic reading (Madisson and Ventsel 2020) of the narrative and aesthetic choices of online conspiracy users (objectives 2, 3, 4).  In order to perform appropriate discourse analysis, the teams will seek to understand the economic, technological, historical and political forces that shape these conspiracy theories and their reception (objective 5). These methods will allow us to identify regional conspiracy vernaculars and concerns, and to link them to aesthetic and narrative strategies that are likely to combine local and global elements, as well as social and infrastructural design. Moreover, we will put particular emphasis on studying cases where conspiracy theorists, individually or collectively, have taken their protest to the streets to understand how the exchange in digital conspiracy theory communities facilitates such actions (objective 6). Comparative analysis, aided by digital methods queries on the combined datasets, will ascertain where concerns converge and diverge, and how conspiracy theories from one region flow to and are adapted within others (objective 4).  Each WP will conduct interviews and engage in observational ethnography with various national and regional fact-checking, disinformation monitoring and anti-extremism civil society organisations, think tanks and regulatory bodies that are interested in conspiracy theories to understand the assumptions about digital conspiracy theories that underscore each organisation’s activities. Taking advice from the European Digital Media Observatory (EDMO), we will interview a variety of organisations, including state run, privately funded, and even lone, small-scale operations, to see how different stakeholders, funding sources  and mission statements might affect assumptions about conspiracy theories and proposed solutions. REDACT has already selected some key example organisations, including Faktograf.hr in Croatia, Centrum Cyfrowe in Poland, Delfi’s fact-checking branch in Estonia, the Berghof Foundation in Germany, and ISD Global in the UK, to work with as CPs. We are particularly interested in addressing the assumptions that underlie automated processes of deplatforming and fact-checking to assess whether these methods can curb the potential harms caused by conspiracy theories (objective 7).  Using our findings, we will be able to improve ways of tackling problematic knowledges in the digital information ecosystem (objective 7). Our overall aim is to develop a comparative perspective to understand how digital conspiracy theories fit into local and transnational spheres of contested knowledge and truth.  Because it is both collecting data from digital social media and conducting ethnographic work within factchecking organisations, we will implement a robust ethical and data management plan. All REDACT research will abide by data protection legislation including but not limited to the Data Protection Act 2018, the UK General Data Protection Regulation, EU General Data Protection Regulation, and any other relevant data protection legislation. Prior to the start of the research, the Research Data Management team at King’s College London (KCL) will produce a data management agreement which will be signed by all REDACT researchers prior to the transfer of personal data to cover its transfer as applicable to the project in compliance with Data Protection Legislation.  KCL has resilient, robust and secure data storage facilities. As well as having standard storage on Sharepoint (up to 25TB), it has recently invested in a new research data storage platform to facilitate projects which generate large research data sets. This new platform provides 2PB of resilient and robust storage capacity, and the platform is specifically designed to cope with complex research datasets. Data type 1, 2, 3 and 4, collected by the UK team, will be stored on either Sharepoint or this new platform. Each regional WP will store data types 1 and 2 on the regional institution’s secure data storage platforms.  With regards to data type 2 (audio recordings and transcripts of interviews), the 5 different regional teams will conduct c. 50 semi-structured interviews in-person or using secure online video conferencing software, such as Microsoft Teams, with people who work at organisations tasked with combatting disinformation.  The data from this part of the research will be processed in a way that is compliant with data protection laws. The audio files will be uploaded to a secure server at the respective institutions, before being deleted from the researchers' laptops, and the team will transcribe the interviews using Nvivo transcription software. Both the audio and transcription files will be encrypted. Transcripts will be pseudonymised if interviewees wish so. | |
|  | What data documentation and metadata you will develop and provide that are accompanying the data? (In documentation provide all information needed for users to be able to read and interpret the data in the future e. g. code books, ReadMe files, etc.) | The design and development of the project website (which will provide a public-facing overview of the project and its findings, including the data visualisations) will be undertaken by a website design team, approved by KCL, working in close coordination with the project team. The design team will be selected on the basis of having a substantial track record of designing, implementing and supporting research projects and producing research websites. The design team will provide training on maintaining the website and assistance in doing so when required. Once the project has ended, data that supports published research and/or has long term value will be deposited with KCL’s Open Research Data System (KORDS) to ensure long term preservation and accessibility. KCL is committed to preserving research data for a minimum of 10 years since last use of the data.  The repository’s standard structured metadata and use of persistent identifiers, such as DOIs and ORCIDs, meets FAIR principles, making data and metadata Findable, Accessible, Interoperable and Reusable. Relevant datasets will be issued a DOI, providing a stable link for access and citation, and this link will be included in data availability statements in all related publications. These datasets will operate under an Open Data Commons Attribution License (ODC-By).  Moreover, proprietary file formats used for data collections will be, when possible, migrated to open source formats to facilitate interoperability across platforms. However, because of the sensitive nature of the material collected (not only might some of the social media data be personally identifiable, but when collated, it could have a potentially radicalising effect if accessed by non-researchers), we will implement strict procedures and regulations that set conditions for how the data can be accessed and used. We will achieve this by asking users to sign an End User Agreement. The datasets will be retained for ten years, after which they will be deleted. | |
| 2. | Ethics, legal and security issues | | |
|  | Are you restricted by a confidentiality agreement? Do you have the necessary permission to obtain process, preserve and share the data? Have the people whose data is being preserved been informed or did they give their consent? What methods will you use to ensure the protection of sensitive data (GDPR special category personal data, specify methods of data anonymization)? | These issues will be managed and obtained through:  - administering informed consent form that will be signed before the start of each interview by selected interviewees. Furthermore, in the process of application we already obtained the endorsement of the organization Faktograf and their general agreement to be involved in the project, as well as their fact-checking journalists as potential interviewees. Such endorsement is important in a sense that Faktograf is still the only fact-checking organization in Croatia. Informed consent and support will be obtained from each interviewee and fact-checking organizations from Bosnia and Herzegovina and Serbia.  - All REDACT research will abide by data protection legislation including but not limited to the Data Protection Act 2018,  the UK General Data Protection Regulation, EU General Data Protection Regulation, and any other relevant data protection legislation. Prior to the start of the research, the Research Data Management team at King’s College London (KCL) will produce a data management agreement which will be signed by all REDACT researchers prior to the transfer of personal data to cover its transfer as applicable to the project in compliance with Data Protection Legislation.  - Data type 1, 2, 3 and 4, collected by the UK team, will be stored on either Sharepoint or this new platform.  Each regional WP will store data types 1 and 2 on the regional institution’s secure data storage platforms. With regards to data type 2 (audio recordings and transcripts of interviews), the 5 different regional teams will conduct c. 50 semi-structured interviews in-person or using secure online video conferencing software, such as Microsoft Teams, with people who work  at organisations tasked with combatting disinformation. The data from this part of the research will be processed in a way that is compliant with data protection laws. The audio files will be uploaded to a secure server at the respective institutions, before being deleted from the researchers' laptops, and the team will transcribe the interviews using Nvivo transcription  software. Both the audio and transcription files will be encrypted. Transcripts will be pseudonymised if interviewees wish so. | |
|  | How will you regulate access to the data and their security? What potential risks do you have to take in consideration? How will you ensure safe sensitive data storage? | Data that supports published research and/or has long term value will be deposited with KCL’s Open Research Data System (KORDS) to ensure long term preservation and accessibility. KCL is committed to preserving research data for a minimum of 10 years since last use of the data. The repository’s standard structured metadata and use of persistent identifiers, such as DOIs and ORCIDs, meets FAIR principles, making data and metadata Findable, Accessible, Interoperable and Reusable. Relevant datasets will be issued a DOI, providing a stable link for access and citation, and this link will be included in data availability statements in all related publications. These datasets will operate under an Open Data Commons Attribution License (ODC-By). Moreover, proprietary file formats used for data collections will be, when possible, migrated to open source formats to facilitate interoperability across platforms. However, because of the sensitive nature of the material collected (not only might some of the social media data be personally identifiable, but when collated, it could have a potentially radicalising effect if accessed by non-researchers), we will implement strict procedures and regulations that set conditions for how the data can be accessed and used. We will achieve this by asking users to sign an End User Agreement. The datasets will be retained for ten years, after which they will be deleted. | |
|  | How will you manage copyright and Intellectual Property Rights issues? Who will be the owner of the data? Which licenses will be applied to the data? What restrictions apply to the reuse of third-party data? | * Each project partner will manage copyright and IPR according to usual institutional practices and regulations * Owners of the data are institutions (universities) involved in the project as research partners. * See explanations above | |
| 3. | Data storage and preservation | | |
|  | How will you store different versions of data during the project?  How will your data be backed-up during the project?  What amount of data are you expecting to be collected and stored during the project (specify in MB/GB/TB) | KCL has resilient, robust and secure data storage facilities. As well as having standard storage on Sharepoint (up to 25TB), it has recently invested in a new research data storage platform to facilitate projects which generate large research data sets. This new platform provides 2PB of resilient and robust storage capacity, and the platform is specifically designed to cope with complex research datasets. Data type 1, 2, 3 and 4, collected by the UK team, will be stored on either Sharepoint or this new platform. Each regional WP will store data types 1 and 2 on the regional institution’s secure data storage platforms. With regards to data type 2 (audio recordings and transcripts of interviews), the 5 different regional teams  will conduct c. 50 semi-structured interviews in-person or using secure online video conferencing software, such as Microsoft Teams, with people who work at organisations tasked with combatting disinformation. The data from this part of the research will be processed in a way that is compliant with data protection laws. The audio files will be uploaded to a secure server at the respective institutions, before being deleted from the researchers' laptops, and the team will transcribe the interviews using Nvivo transcription software. Both the audio and transcription files will be encrypted. Transcripts will be pseudonymised if interviewees wish so.  The amount of four main data types we expect are:  (1) a selection of thematically significant text, image and audio-visual materials collected from websites and social media platforms, to be used as the basis for the team’s discourse analysis (stored as RTF and PDF files for text and image, with links to audio-visual materials as necessary, c. 200GB);  (2) audio recordings and transcripts of interviews with organisations tasked with combatting disinformation  (c. 50GB) (audio stored in MP3 format and transcribed using Nvivo software, c. 100GB);  (3) big data searchable archives of individual social media platforms, collected using custom scraping tools (see below, c. 10TB); and  (4) data visualisations resulting from those big data queries (created using Gephi, c. 50GB). | |
|  | How will your dataset be curated and preserved during the project and after the project?  What file formats will be used for data storage?  What amount of data are you expecting to be collected and stored after the project (specify in MB/GB/TB) | Datatype 1, 2, 3 and 4, collected by the UK team, will be stored on either Sharepoint or this new platform. Each regional WP will store data types 1 and 2 on the regional institution’s secure data storage platforms. With regards to data type 2 (audio recordings and transcripts of interviews), the 5 different regional teams will conduct c. 50 semi-structured interviews in-person or using secure online video conferencing software, such as Microsoft Teams, with people who work at organisations tasked with combatting disinformation. The data from this part of the research will be processed in a way that is compliant with data protection laws. The audio files will be uploaded to a secure server at the respective institutions, before being deleted from the researchers' laptops, and the team will transcribe the interviews using Nvivo transcription software. Both the audio and transcription files will be encrypted.  The amount of four main data types we expect are:  (1) a selection of thematically significant text, image and audio-visual materials collected from websites and social media platforms, to be used as the basis for the team’s discourse analysis (stored as RTF and PDF files for text and image, with links to audio-visual materials as necessary, c. 200GB);  (2) audio recordings and transcripts of interviews with organisations tasked with combatting disinformation (c. 50GB) (audio stored in MP3 format and transcribed using Nvivo software, c. 100GB);  (3) big data searchable archives of individual social media platforms, collected using custom scraping tools (see below, c. 10TB); and  (4) data visualisations resulting from those big data queries (created using Gephi, c. 50GB). | |
| 4. | Data sharing and reuse | | |
|  | How and where will the data be shared? On which repository do you plan to share your data? How will potential users find out about your data? | Data will be shared at the Sharepoint, as described in the previous sections. | |
|  | If there is any data which cannot be shared (due to legal, ethical, copyright, confidentiality reasons) explain the reasons of restrictions | No | |
|  | Confirm that the digital repository you choose is in line with the FAIR principles | I confirm! | |
|  | Please confirm that you will use a digital repository maintained by a non-profit organisation (if not please explain why) | I confirm! | |