CLEANUP:
Machine Learning for the Anonymisation of Unstructured Personal Data


UiO : Department of Informatics
University of Oslo

NTNU
Norwegian University of Science and Technology


UNIVERSITAT ROVIRA i VIRGILI


UiO : Faculty of Law
University of Oslo


ARKIVVERKET

1 Excellence

1.1 State of the art, knowledge needs and project objectives

Public and private organisations are increasingly aware of the value of the data they collect or produce. Access to high-quality data is often a decisive factor in fostering scientific and technological innovations. However, data that may reveal personal information about individuals must also comply with existing privacy and data protection laws, such as the *General Data Protection Regulation* (GDPR) newly introduced in Europe. In particular, personal data collected for one specific purpose cannot be repurposed for secondary uses (such as quality assurance or scientific research) without proper legal ground such as the explicit consent of the individuals to whom the data refers.

One solution is to use *de-identification* and *anonymisation* techniques[1] to protect the privacy of registered individuals. When the data is structured (i.e. when it can be expressed as a table with fixed attributes, each with a predefined range of possible values), anonymisation can be enforced through well-established techniques such as k-anonymity or differential privacy. However, these methods were originally designed for tabular data and are difficult to apply to unstructured formats such as text documents. The lack of suitable anonymisation techniques for these formats is an important problem for many organisations, as unstructured or semi-structured data often make up a large part of modern data warehouses – by some estimates, up to 80% of the world's data is unstructured (Rizkallah, 2017). This technological gap may also hinder scientific advances in social sciences, law, psychology and medicine (since most medical information is only available in text records).

The CLEANUP project aims to develop novel *computational models* and *processing techniques* to automatically anonymise unstructured data, with a focus on text documents. The project will also design dedicated *evaluation methods* to assess the empirical performance of text anonymisation mechanisms, and examine how these metrics should be interpreted from a legal perspective. Finally, the project will investigate how these technological solutions can be integrated into organisational processes – in particular how *quality control* can be effectively performed, and how the level and type of anonymisation can be *parametrised* to meet the specific needs of the data owner.

To achieve these objectives, the project brings together a consortium of researchers with expertise in machine learning, natural language processing, computational privacy, statistical modelling, health informatics and IT law. In addition, external partners from the public and private sector will also contribute to the research objectives with their data, use cases and domain knowledge.

[1] Anonymisation is often distinguished from de-identification by requiring the process to be *irreversible*, without any technical means of mapping back pseudonyms or coded values to the original identifiers. In this proposal, the term "anonymisation" will be employed in a broad sense to comprise any process seeking to modify a dataset to prevent the disclosure of personal information (while preserving as much semantic content as possible).

1.2 Novelty and ambition

Background

The development of automated methods for text anonymisation has so far taken place in two research fields with little interaction with one another: *natural language processing* (NLP) on one hand and *privacy-preserving data publishing* (PPDP) on the other.

NLP approaches to text anonymisation have mostly focused on the detection of personal identifiers in medical health records (Jegga et al., 2013). These identifiers include direct identifiers such as person names or national ID numbers along with more indirect identifiers (often called *quasi-identifiers*) such as dates or place names[2]. Early approaches relied on handcrafted rules that search for the occurrence of specific patterns in the document (Neamatullah et al., 2008). Alternatively, one can frame the detection of personal identifiers as a sequence labelling problem where the goal is to decide for each token whether it should be anonymised or not – and if the answer is positive, what type of identifier it belongs to. Based on documents manually annotated by human experts, the detection of identifiers can be learned using machine learning models such as conditional random fields (Aberdeen et al., 2010) or recurrent neural networks (Dernoncourt et al., 2017).

Although these NLP approaches are undeniably useful, they suffer from a number of shortcomings. One limitation is that these methods typically erase or mask all detected identifiers without regard to their actual disclosure risk. For instance, mentions of the patient's profession in a medical record will be systematically removed regardless of whether the combination of this variable with other quasi-identifiers leads to an increased disclosure risk – and without taking into account the (medical) importance of this information. Furthermore, these NLP approaches are limited to well-defined categories such as names, places, and dates (which are relatively easy to annotate and detect) and do not consider how less conspicuous elements (such as the mention of a particular event in which the person was involved) may also play a role in re-identifying the individual.

The notion of disclosure risk lies, on the other hand, at the centre of anonymisation approaches based on privacy-preserving data publishing (PPDP). These approaches typically operate by generalising sensitive words – for instance replacing the term "surgeon" with the more abstract "doctor". The k-safety and k-confusability models (Chakaravarthy et al., 2008; Cumby and Ghani, 2011) extend the k-anonymity model used for tabular data and require each sensitive entity present in a document collection to be indistinguishable from at least $k-1$ other entities in that collection. The t-plausibility model (Anandan et al., 2012) generalises sensitive words such that at least t distinct documents can be derived from the protected document by re-specialising the generalised words. Finally, Sánchez and Batet (2016) present a privacy model for document sanitisation relying on the mutual information between the entities to protect and the words occurring in the document. Words that, individually or in aggregate, may disclose the entities to protect are then generalised.

The strength of PPDP-based approaches is their grounding in data semantics and information theory, making it possible to reason over disclosure risks in a principled manner. However, these approaches essentially represent documents as bags of terms void of internal structure and ignore the multiple types of linguistic dependencies between the terms occurring in a document. For instance, PPDP-based approaches would typically sanitise the two sentences "*Peter lives in Oslo*" and "*Peter never lived in Oslo*" in the same manner, ignoring the negation marker in the second sentence. Furthermore, these approaches are unable to determine whether a sensitive term is linguistically linked to a particular individual or not. This may lead to spurious modifications of words that are not in any way associated with the persons mentioned in the document. A court ruling might for instance contain paragraphs with generic legal interpretations that are not tied to a personal situation and should as a consequence be ignored from the anonymisation.

Finally, a powerful alternative to defining privacy through properties of the data, e.g. containing "personal information" or not, is by defining privacy in terms of randomised algorithms for comput-

[2]Quasi-identifiers do not typically disclose the identity of registered individuals in isolation, but may do so when combined with one another. For instance, the combination of gender, birth date and postal code can be used to identify between 63 and 87% of the U.S. population (Golle, 2006).

ing statistics from the data. An emerging standard based on this idea is differential privacy (Dwork et al., 2006). Differential privacy provides strong guarantees that the computed statistics cannot be used to learn anything substantial about any individual in the data, while simultaneously allowing accurate statistics to be computed. However, the use of differential privacy on text data has so far been the object of few studies, typically limited to simple processing techniques. Fernandes et al. (2018) describe for instance a method to convert the original bag-of-words of a document into a "noisy" bag-of-words using generalised differential privacy.

Project objectives

The central objective of the CLEANUP project is to bring together the NLP and PPDP approaches to text anonymisation, which have until now unfolded in relative isolation from one another. In particular, the project aims to develop *a new generation of anonymisation models* that simultaneously:

1. Take advantage of state-of-the-art language processing techniques to derive fine-grained records of the individuals referred to in a given document;
2. Connect these individual records to principled measures of disclosure risk and data utility, with the goal of modifying the document in a way that prevents identity and/or attribute disclosure while retaining as much as possible the document's semantic content.

The anonymisation quality resulting from these techniques will be assessed with a battery of novel *evaluation metrics* relying on both intrinsic and extrinsic factors. The participation of external partners coming from various branches of the public and private sector will also allow the CLEANUP project to validate the anonymisation methods across a broad range of document types.

The CLEANUP project will also include legal perspectives on the anonymisation of text data. In particular, the project will seek to connect empirical measures of disclosure risk with legal regulations and investigate how privacy risk assessments are to be performed when dealing with large amounts of text data. The project will also analyse how trade-offs between privacy concerns and other legitimate interests can be formalised and integrated into machine learning models.

Finally, another distinctive feature of the project is its focus on Norwegian, which is a language with relatively limited language technology resources. The project is therefore expected to contribute to the development of new linguistic resources and software tools for Norwegian.

1.3 Research questions and hypotheses, theoretical approach and methodology

The two central research questions addressed by the CLEANUP project are:

1. Can we develop text anonymisation methods that can take into account both the linguistic structure of the document and the disclosure risks associated with each individual to protect?
2. Do these methods improve upon the state-of-the-art? Do their results transfer across domains, and how well are they aligned with existing regulations on privacy and data protection?

To address these overarching questions, the project will work on the following research topics:

Detection of personal identifiers: The first step in the anonymisation process is to detect the occurrences of direct and indirect identifiers in the document. To this end, the project will explore the use of neural architectures for sequence labelling. To mitigate the scarcity of annotated data, the project will make use of pre-training approaches that have been shown to obtain state-of-the-art results on a wide array of NLP tasks (Devlin et al., 2018). The resulting models will then be fine-tuned using domain-specific data collected for the project. The project will also investigate the use of techniques based on Open Information Extraction (Niklaus et al., 2018) to extract additional content that are not quasi-identifiers in a traditional sense but may nevertheless increase the re-identification risk through semantic inferences.

Extraction of individual records: Text documents may refer to several individuals whose identity must be protected. For instance, court rulings may include multiple defendants, plaintiffs and witnesses. In such cases, the anonymisation model must group the identifiers into individual records that aggregates all pieces of information about a given person, and apply co-reference resolution to group together mentions referring to the same person (Lee et al., 2017). In contrast to structured datasets (where each value is always unique and known with certainty), these individual records may include uncertain values, notably due to linguistic ambiguities[3].

Generation of synthetic data: The use of real-world personal data for R&D purposes is often problematic due to privacy and data protection issues. One alternative is to rely on *synthetic data* (Velupillai et al., 2018). Although generation methods for structured data are well established, the generation of synthetic documents has only recently been studied, notably using encoder-decoder architectures (Lee, 2018). However, these approaches are intended for short documents and are poorly suited to longer texts due their limited account of long-range dependencies. An alternative strategy, which will be explored in this project, is to generate synthetic data through sequences of transformations based on an initial set of documents (Guu et al., 2018).

Algorithms for text sanitisation: Based on these records, the documents' content must be modified ("sanitised") to prevent re-identification and/or attribute disclosure. This can be framed as a constrained optimisation problem, where the goal is to minimise the disclosure risk while retaining as much semantic content as possible. The project will extend existing approaches by taking into account the syntactic and semantic structure of the text as well as document-level constraints (i.e. the same entity should be edited consistently through a document). As the individual records extracted from text may include uncertain values, the sanitisation algorithms will also need to operate on probabilistic representations. This could be achieved by using optimisation methods under uncertainty such as stochastic programming.

Differential privacy on text databases: Differentially private methods often require structured inputs, making them challenging to apply to text data. Differentially private deep learning models have recently been developed, including recurrent architectures suitable for texts (McMahan et al., 2017), but training those networks requires substantial amounts of data. The project will investigate when differentially private deep learning (e.g. by sampling from generative models) are advantageous wrt. less expressive (but potentially less data intensive) methods on top of text representations such as word and document embeddings.

Ontologies for entity generalisation: An important strategy to sanitise words or phrases that are deemed too sensitive is *generalisation*, i.e. replacing a specific term by a more abstract one. The project will construct ontologies specifically designed for the purpose of text sanitisation, based on existing resources such as DBpedia (Lehmann et al., 2015) or WordNet databases.

Evaluation methods: Existing text anonymisation techniques rely on simple evaluation methods, such as measuring the precision, recall and F_1 based on documents that are manually anonymised by human experts. One shortcoming of these methods is their exclusive focus on the detection step, whereas text sanitisation is often left out. In addition, these methods ignore the fact that sensitive terms are not equally important in terms of disclosure risk. The project will design new evaluation methods that reflect these characteristics. In addition, the project will investigate how to connect these empirical measures of disclosure risk with legal regulations on privacy and data protection, in particular with respect to privacy risk assessments (Hintze, 2017).

Interactive quality control: In most cases, anonymisation models are not meant to operate fully automatically, given the real-world consequences of false negatives on the privacy of registered individuals. Rather, these models are best applied in close interaction with humans experts through a user interface. In this interactive setting, the anonymisation model should not only

[3]For instance, in the sentence "*The defendant, Mr. John Doe, allegedly assaulted Mr. Peter when he arrived to his home address in Southwark*", the identifier "*home address in Southwark*" may be connected to Mr. Doe or Mr. Peter.

provide editing suggestions on sensitive terms but also explanations on *why* these terms should be edited and *what kind/level of disclosure risk* may occur. How to make the predictions of deep neural networks transparent and interpretable is notoriously difficult, but some methods do exist (Lundberg and Lee, 2017) and their use will be explored in the project[4].

Model parametrisation: Anonymisation techniques must strike a delicate balance between the level of protection and the preservation of the data content. This balance may vary depending on multiple factors, in particular legal/regulatory constraints and the intended target audience of the protected data. The project will look at how anonymisation techniques can be made *adaptive* (using tunable meta-parameters) to the particular needs of the application domain.

In close collaboration with the external partners, the project will develop an *annotation scheme* to enrich the data with additional information about the textual entries that should be sanitised. A fraction of the provided data will then be annotated using this scheme. Based on these collected datasets, the project will design algorithms to generate large amounts of synthetic data (as described above), which will be in turn employed to train text anonymisation models.

Risk mitigation and ethical issues

The most salient risk for the CLEANUP project relates to the data collection process, since the generation of synthetic data is dependent on the availability of an initial set of documents. This risk will be mitigated by a combination of several measures. The most important is the participation of several external partners, each owning large databases of unstructured personal data. Furthermore, the patient records stored at the *Norwegian Health Archives* correspond to records of individuals deceased for at least 10 years. As existing legal regulations on privacy and data protection only apply to living individuals, there is no need for a legal ground (such as consent) to collect the data, although an application to the *Regional Committees for Medical and Health Research Ethics* is still required due to the sensitive nature of the data. Finally, some international datasets of textual records are also available, such as the annotated patient records collected by Yang and Garibaldi (2015).

The project intends to take privacy concerns very seriously and will adopt a range of security-enhancing measures, such as storing the project data on secure servers with strict access control and providing a training course on data privacy to all researchers involved in the project. Data Protection Impact Assessments (DPIA) will also be undertaken in close collaboration with the external partners and their Data Protection Officers. These legal procedures will be launched as soon as project starts to ensure the data can be made available to the project researchers in due course.

2 Impact

2.1 Potential impact of the proposed approach

Many organisations struggle to manage the personal data they gather or produce as part of their activities. A large part of this data takes the form of text documents, which may contain information relating to customers, patients, welfare recipients, or even defendants in court cases. This data is often highly valuable, both for the organisation that owns it and for society at large. For instance, customer data may be used to improve the company's services and user experience. Similarly, patient records are essential for biomedical research, and court cases constitute a key resource for legal professions. However, much of this data includes personal information collected for a particular purpose, such as providing healthcare services. According to privacy regulations such as the European *General Data Protection Regulation* (GDPR), such data cannot be used for other, secondary purposes – and even less be shared with third parties – without the affirmative consent of each individual.

De-identification and anonymisation are therefore essential to allow these organisations to take full advantage of their data without compromising the privacy of registered individuals. Manual

[4]This is also an active research topic in BigInsight, for which the Norwegian Computing Center is host institution.

anonymisation is, however, extremely costly and prone to human errors, omissions and inconsistencies[5]. The development of automated methods is thus expected to provide substantial benefits to public and private organisations in charge of processing personal data in unstructured forms, as is the case for the six external partners to this proposal. The outcomes of the project may also serve as an important enabling technology in biomedical research, as privacy concerns are a major impediment to data sharing and "open science" efforts in this field (Velupillai et al., 2018).

Finally, the project is also expected to impact the development of language resources for Norwegian, which remain relatively scarce compared to other languages.

2.2 Measures for communication and exploitation

Dissemination efforts will target a range of complementary channels. As is common practice in computer science, scholarly dissemination will follow a process of staged publication, starting with early results in specialised workshops, followed up by consolidated experiments in selective conferences and finally exposed in journal articles. The project results will be presented in established international venues such as ACL, EMNLP, ARES or ICTAI along with top-tiered journals such as *Information Sciences*, *Knowledge-Based Systems*, *Computational Linguistics*, *Journal of Biomedical Informatics*, *Artificial Intelligence in Medicine*, *Artificial Intelligence and Law* and *Harvard Journal of Law and Technology*. To facilitate a wide circulation of the project results, the software tools developed during the project will be released under an open-source licence and made available on a code repository.

Thanks to the participation of five external partners, the project will be in regular contact with its "stakeholders", who will be invited to yearly project meetings and other project-related events. To facilitate the transfer of research results into practical solutions, these external partners will be regularly invited to test software prototypes and provide feedback to the research team. Several partners have already indicated their interest in implementing the anonymisation methods developed by the project into their own IT infrastructure. Furthermore, the *Language Technology Group* also participates in the BigMed project, which is a large Norwegian research effort on data-driven precision medicine in which the anonymisation of patient records is an important challenge.

Popular dissemination will also play an important role through the project. The project partners have a strong track record in public outreach activities. The interactions between artificial intelligence, health and privacy is currently attracting much attention from the media, and informing the public opinion about the promises and limitations of anonymisation techniques is therefore essential.

Finally, the project will also take advantage of its international partners to actively promote the project beyond Norwegian borders, as the challenges addressed by the project are international in nature. A subsequent scale-up through a EU project proposal will also be considered.

3 Implementation

3.1 Project manager and project group

The project consortium is composed of five research institutions:

- **NR** (Norwegian Computing Centre) is one of Norway's leading research institutions within statistical modelling, machine learning and ICT. NR currently leads the BigInsight centre of excellence for research-based innovation, which aims to produce innovative solutions for the knowledge economy through novel statistical and machine learning methodologies.

  Pierre Lison is a senior research scientist at NR and will lead the CLEANUP project. He received in 2014 his PhD in computer science from the University of Oslo. Pierre has extensive research experience in natural language processing and machine learning and has led several R&D projects in these fields, including two recent applied projects on the de-identification of speech and text data. He is also a member of the Young Academy of Norway.

[5] Neamatullah et al. (2008) provide evidence of such inconsistencies in the de-identification of patient records.

Anders Løland is assistant research director and research leader at NR, where he has worked since 2001. He holds a Dr.Philos. from the Department of Mathematics, University of Oslo. Anders currenctly leads the "AI – Explanation & Law" project in BigInsight, which deals with interpretable and transparent predictions from machine learning models.

- **LTG** (Language Technology Group, University of Oslo) has broad expertise in the areas of natural language processing and machine learning and conducts research in areas such as semantic vector modelling, syntactic and semantic parsing, sentiment analysis, and clinical NLP.

  Lilja Øvrelid is associate professor at the Department of Informatics and LTG's current group leader. Øvrelid leads the LTG involvement in the BigMed project on ICT support for precision medicine and is also co–project manager for the IKTPLUSS project on Sentiment Analysis for Norwegian Text (SANT). She has been involved in the creation of several resources and tools for the analysis of Norwegian texts and has published extensively on machine learning applications for NLP.

- **IIK** (Department of Information Security and Communication Technology, Norwegian University of Science and Technology – NTNU) hosts the Center for Cyber and Information Security (CCIS) with 25 participating partner institutions, including governmental departments and national infrastructure operators. IIK also leads the National Research School on Information Security (COINS) that integrates Norwegian research groups in information security.

  Staal Vinterbo is a Professor in IIK at NTNU. He is a computer scientist and has during the last two decades been researching how to use population data under privacy constraints. He has been invited multiple times to serve as a privacy expert in the US national discourse on health information privacy, including at events hosted by the US Dept. of Health and Human Sciences and the Centers for Disease Control. He has also served as a director of research and development in "Integrating Data for Analysis, Anonymization and SHaring" (iDASH), one of the US National Centers for Biomedical Computing funded by the NIH Roadmap initiative.

- **DPIL** (Department of Public and International Law, University of Oslo) has many leading Nordic legal scholars in law and technology studies, including on questions of digitalisation, legal technology, data access and ownership, privacy and consent. The Department notably hosts the Digital Lawyer project (led by Prof. Malcolm Langford) and participates to the BigMed project on legal issues (Ph.D Fellow Anne Kjersti Befring co-coordinates the work package on law and ethics).

  Malcolm Langford is a Professor of Public Law, University of Oslo, and Co-Director of the Centre on Law and Social Transformation, CMI and University of Bergen. He leads the Digital Lawyer project at the Faculty of Law, coordinates the Faculty's working group on Law and Technology, and is the course coordinator for the new subject Legal Technology: Artificial Intelligence and the Law. A trained lawyer and economist, he leads and coordinates a number of large externally-funded research projects on empirical legal studies and computational legal studies and is the Co-Editor of the Cambridge University Book Series on Globalization and Human Rights.

- **CRISES** (Security & Privacy Group, Universitat Rovira i Virgili, Spain) is the strongest research group in data privacy in Spain and one of the top groups at European level. It holds the UNESCO Chair in Data Privacy and leads the CYBERCAT-Center for Cybersecurity Research of Catalonia.

  David Sánchez is an Associate Professor at the Universitat Rovira i Virgili. He received in 2008 his PhD in computer science from the Polytechnic University of Catalonia. He has authored more than 150 papers and has participated in 7 EU projects (1 as coordinator). He has an extensive background on knowledge-based systems, semantics and data privacy.

  Montserrat Batet is a senior researcher at the Open University of Catalonia. She received in 2011 her PhD in computer science from the Universitat Rovira i Virgili. She has authored around 60 papers on data classification, semantic similarity and data privacy.

  Josep Domingo-Ferrer is a Distinguished Professor at the Universitat Rovira i Virgili. He received in 1991 his PhD in computer science from the Autonomous University of Barcelona. He is an ACM Distinguished Scientist and Fellow of IEEE. He served as coordinator for several large EU-funded projects and authored more than 400 publications on data privacy.

The project also includes *five external partner organisations* who will contribute to the project objectives with their data collections, use cases and domain expertise:

- **NAV** (Norwegian Labour and Welfare Administration) administers a third of the national budget, employs 19 000 people, and services almost 2.8 million people through a broad range of public benefit and social service schemes. NAV manages and retains stewardship of several large data sources on its users and services, much of which in unstructured text. NAV IT Data & Insight employs these data to innovate and improve its services, contingent upon strict data privacy practices. This project extends ongoing efforts in privacy-preserving data analysis in NAV.
- **NHA** (Norwegian Health Archives) is mandated to receive and preserve patient records from public and private hospitals, with the goal of disseminating health resources to researchers and next-of-kin in compliance with regulations on privacy and data protection. Patient records archived at NHA must relate to individuals that have been deceased for at least 10 years. The NHA is currently being established at Tynset with 50 employees, and is part of the National Archives of Norway.
- **Gjensidige** (ASA) is Norway's largest insurer with more than four thousand employees, business in six countries and a written premium of more than 21 billion NOK. Gjensidige focuses on running efficient operations supported by a market-leading digital customer experience and data analytics embedded in the entire value chain among its strategic targets.
- **Lovdata** is the oldest and largest source of digital legal information in Norway and is a pioneer in the field of legal informatics. Lovdata has over a decade of experience with semi-automatic anonymisation of personal information in legal documents.
- **USIT** (University Center for Information Technology, University of Oslo) employs over 200 technical staff in support of the university's research and teaching activities. It provides IT services in the form of software, computational and storage resources and access to data collections. It recently launched TSD, a national platform to collect, store and analyse sensitive data.
- **DNB** (ASA) is Norway's largest financial services group and one of the largest in the Nordic region in terms of market capitalisation. DNB offers a full range of financial services, including loans, savings, advisory services, insurance and pension products for retail and corporate customers.

Finally, the project also includes *two affiliated researchers* with whom the project will collaborate closely (through e.g. the organisation of common seminars or workshops):

- *Sumithra Velupillai* is a lecturer in Applied Health Informatics at the Institute of Psychiatry, Psychology and Neuroscience (IoPPN) at King's College London and NIHR Maudsley Biomedical Research Centre. She received her PhD in 2012 in Computer Science from Stockholm University, and has held postdoc positions in the US (UCSD, University of Utah) and an International Career Grant at KTH/King's College London. Her research interests focus on information extraction applied to clinical texts and has worked on both Swedish and English data.
- *Christos Dimitrakakis* is an associate professor at the Department of Informatics, University of Oslo, and also holds a senior researcher position at Chalmers university of technology, Sweden. He received his PhD in 2006 from the Swiss Federal Institute of Technology (EPFL). His research interests include decision theory and reinforcement learning, security and privacy, and he has authored more than 60 publications in these fields.

3.2 Project organisation and management

The Norwegian Computing Centre (NR) will be responsible for the overall management of the project. The project is planned for a duration of four years and is divided in seven work packages. The work packages are illustrated in Figure 1. The project will first collect data together with the external partners and annotate part of this data to mark sensitive entries (part of the project budget is specifically devoted to this effort). Synthetic data will be generated on that basis and employed in WP2-WP5.

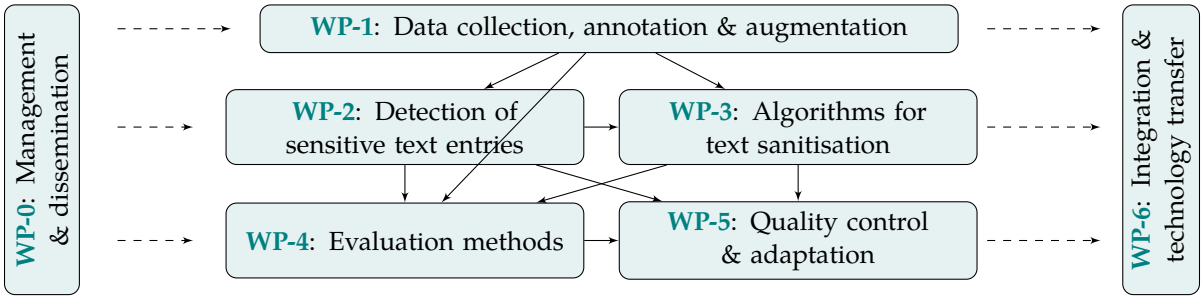

Figure 1: Organisation of work packages.

WP-0: Project management and dissemination	**Lead:** NR	**Period**: M1-M48
Task 0.1: Project management (NR). Coordination of research activities, administration and internal communication, organisation of project-related events, reporting, quality assurance. **Task 0.2: Dissemination** (all). Research publications, participation in national and international conferences, public outreach activities, demonstration of prototypes to stakeholders, etc.		

WP-1: Data collection, annotation & augmentation	**Lead:** NR	**Period**: M1-M21
Task 1.1: Data collection and storage (NR). Coordination with external partners and data protection authorities to collect the data necessary to the project. Storage on secure server. **Task 1.2: Data annotation** (NR,all). Development of a common annotation scheme to mark textual entries to be sanitised. Coordination of the annotation effort among research partners. **Task 1.3: Generation of synthetic data** (NR,LTG,IIK,NAV). New methods for producing synthetic data based on existing documents through e.g. sequences of text transformations.		

WP-2: Detection of sensitive text entries	**Lead:** LTG	**Period**: M9-M39
Task 2.1: Detection of personal identifiers (LTG, NR). Pre-training of neural language understanding models for Norwegian and fine-tuning to detect direct and indirect identifiers. **Task 2.2: Extraction of individual records** (NR, LTG). Aggregation of detected identifiers into individual records (based on e.g. co-reference resolution techniques).		

WP-3: Algorithms for text sanitisation	**Lead:** IIK	**Period**: M9-M45
Task 3.1: Ontologies for entity generalisation (LTG, NR). Construction/adaptation of ontologies for the purpose of replacing sensitive text entries with more generic terms or phrases. **Task 3.2: Sanitisation strategies** (CRISES,NR). Text sanitisation with constraints on syntactic or semantic structure (e.g document-level consistency) and ability to deal with uncertain inputs. **Task 3.3: Differentially private text data queries** (IIK). Statistical inference from text databases under differential privacy constraints.		

WP-4: Evaluation methods	**Lead:** CRISES	**Period**: M27-M48
Task 4.1: Evaluation of text anonymisation (CRISES, NR). Design of novel evaluation methods to assess the anonymisation quality and data utility, using both intrinsic and extrinsic factors. **Task 4.2: Legal perspectives** (DPIL). Link between quantitative evaluation measures and legal regulations on privacy and data protection, in particular privacy risk assessments.		

WP-5: Quality control and adaptation	**Lead:** NR	**Period**: M30-M48
Task 5.1: Interactive quality control (NR, DPIL, NAV). Interface to apply anonymisation tool interactively, including explanations of editing suggestions and estimation of disclosure risk.		
Task 5.2: Adaptive anonymisation models (NR, IIK, DPIL). Parametrisation of anonymisation models to control the level and type of anonymisation at runtime.		

WP-6: Integration & Technology transfer	**Period**: M6-M48
Task 6.1: Technology transfer (all) Collaboration between the R&D partners and the external organisations on the transfer of technological solutions from WP2-WP5 in their infrastructure.	
Task 6.2: Integration (external partners) . Experiments with the anonymisation techniques from WP2-WP5 on in-house data and (when applicable) integration in IT infrastructure.	

In addition to various informal meetings between partners, the project will organise yearly meetings/workshops to present ongoing research work and draft future plans. The project partners will be actively encouraged to collaborate across institutions, notably through research visits.

The software developed for the project will be released under an open-source license and published on a public repository. Sensitive datasets will be stored on secure, encrypted servers especially designed for research on sensitive data and managed by one project partner (USIT).

References

Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B., and Hirschman, L. (2010). The MITRE identification scrubber toolkit: design, training, and assessment. *International journal of medical informatics*, 79(12):849–859.

Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P., and Si, L. (2012). t-plausibility: Generalizing words to desensitize text. *Transactions on Data Privacy*, 5(3):505–534.

Chakaravarthy, V. T., Gupta, H., Roy, P., and Mohania, M. K. (2008). Efficient techniques for document sanitization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 843–852. ACM.

Cumby, C. and Ghani, R. (2011). A machine learning based system for semi-automatically redacting documents. In *Proceedings of the Twenty-Third Conference on Innovative Applications of Artificial Intelligence (IAAI 2011)*.

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Conference on Theory of Cryptography*.

Fernandes, N., Dras, M., and McIver, A. (2018). Generalised differential privacy for text document processing. *CoRR*, abs/1811.10256.

Golle, P. (2006). Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 77–80. ACM.

Guu, K., Hashimoto, T. B., Oren, Y., and Liang, P. (2018). Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Hintze, M. (2017). Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency. *International Data Privacy Law*, 8(1):86–101.

Jegga, A., Solti, I., Molnar, K., Marsolo, K., Stoutenborough, L., Deleger, L., Kaiser, M., Li, Q., Lingren, T., Savova, G., and Xia, F. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20(1):84–94.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 188–197.

Lee, S. H. (2018). Natural language generation for electronic health records. *npj Digital Medicine*, 1(1):63.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4765–4774.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. (2017). Learning Differentially Private Recurrent Language Models. *arXiv:1710.06963 [cs]*.

Neamatullah, I., Douglass, M. M., Li-wei, H. L., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):32.

Niklaus, C., Cetto, M., Freitas, A., and Handschuh, S. (2018). A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878.

Rizkallah, J. (2017). The big (unstructured) data problem. *Forbes*. June 5, 2017.

Sánchez, D. and Batet, M. (2016). C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., and Dutta, R. (2018). Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*.

Yang, H. and Garibaldi, J. M. (2015). Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:30 – 38.