Project CODE Data Management Plan

1. Data Summary

1.1 Purpose of Data Collection/Generation

The purpose of data collection and generation is to understand the dynamics of synthetic geolocalized populations, epidemic outbreaks, and the interplay of political discourse on social media platforms, particularly X (formerly Twitter). This data will facilitate machine learning and data mining algorithms for classification, clustering, and pattern discovery. The data will benefit researchers, policy analysts, public health experts, and data journalists, aiding in combating disinformation and enhancing digital literacy.

1.2 Relation to Project Objectives

The data collection supports objectives including:

1. Monitoring socio-political and health-related conversations on social media.
2. Designing tools for analysts, experts, and researchers to assess misinformation.
3. Generating reliable synthetic populations for epidemiological modeling.
4. Understanding information flow dynamics and their spatial covariates.

1.3 Types and Formats of Data

- **Census Data:** Aggregated demographic data in formats such as CSV and JSON.
- **Contact Matrices:** Tabular data in CSV format describing interaction patterns.
- **Synthetic Populations:** Geo-referenced datasets generated using algorithms, stored in formats like CSV or GeoJSON.
- **Epidemic Data:** Epidemiological records and time-series data in CSV and XLSX formats.
- **Social Media Data:** JSON files containing tweets, metadata, and user interaction graphs.

1.4 Re-use of Existing Data

Existing data repositories will be used whenever possible. Examples include:

- **Social Science One Facebook Dataset**: To be reused under strict access agreements.
- Online repositories such as:
    - 40twita 1.0: A collection of Italian Tweets during the COVID-19 Pandemic, http://twita.di.unito.it/dataset/40wita
    - the COVID-19 integrated surveillance data in Italy, https://www.epicentro.iss.it/en/coronavirus/sars-cov-2-dashboard

1.5 Origin of Data

- **Census Data:** National statistics offices.
- **Social Media Data:** X API and secondary sources.
- **Epidemic Data:** WHO, CDC, and academic repositories.

1.6 Expected Data Size

Initial estimates suggest a total of a few terabytes, depending on collection scope and timeframe.

2. FAIR Data

2.1. Making Data Findable

- **Naming Scheme:** E.g., `Twitter.20250101.20251231.COVID19.csv`
- **Metadata Standards:** Use of JSON schemas for social media; Dublin Core for demographic data.
- **Indexing and Searchability:** Hosted on Zenodo and GitHub for public datasets.

2.2. Making Data Openly Accessible

- **Open Access:** Anonymized datasets made available on public repositories.
- **Restricted Access:** Sensitive data, e.g., from Social Science One, will require formal access requests.
- **Access Methods:** Password-protected downloads; APIs with secure tokens.

2.3. Making Data Interoperable

- **Standards:** Use of JSON, CSV, and RDF to ensure compatibility.
- **Vocabulary:** Adoption of domain-specific ontologies like Schema.org or epidemiological standards.

2.4. Increase Data Re-use

- **Licenses:** Open datasets under Creative Commons CC-BY or CC0.
- **Embargo Periods:** None anticipated.
- **Quality Assurance:** Data will undergo validation using cross-validation and anomaly detection techniques.
- **Reusability Duration:** Indefinitely for most datasets.

3. Allocation of Resources

- Public repositories, open source software, and servers already hosted at CNR will be used to cut all possible costs related to data storage, software licenses and infrastructure for high-performance computing.
- Due to the prohibitive costs of premium APIs for data collections following the recent management change of X, we will mostly make use of public APIs or previously collected data.

4. Data Security

- All private and sensitive data will be stored in local, encrypted drives with routine backups.
- Access control will be guaranteed by the standard mechanisms put in place by the project partners (e.g., role-based permissions to restrict access, multi-factor authentication, regular audits of access logs).
- Data will only be transmitted using encrypted channels such as HTTPS and SFTP to prevent unauthorized interception.

5. Ethical and Legal Compliance

- We do not expect to treat any data that raises ethical concerns, but the CNR Research Ethics and Integrity Committee will be consulted where necessary.
- All activities will comply with GDPR, local data protection laws, and terms of service agreements for APIs and datasets.
- Public social media data will be anonymized and processed in compliance with ethical research standards, avoiding any publication of personal identifiers.

6. Data Sharing and Repositories

- Processed data and code, stripped of any private or sensitive information, will be shared on open platforms like Zenodo, Figshare, and GitHub.
- Sensitive datasets will be shared under data-sharing agreements with proper credentialing and justification.
- Comprehensive metadata, user manuals, and analysis scripts will accompany all datasets to ensure reproducibility.

7. Monitoring, Review and Preservation

- The DMP will be reviewed during the project lifespan to accommodate technological, legal, or procedural changes.
- Metrics will include the volume of data processed, accessibility of shared datasets, and compliance with FAIR principles.
- Data will be preserved for a minimum of five years post-project, with extensions as required by stakeholders or funding bodies.

- Secure data erasure protocols will be employed for datasets no longer needed, using certified data destruction tools.

8. Conclusion

This Data Management Plan ensures the integrity, security, and accessibility of data collected and generated, aligning with the goals of the project and best practices in research data management.

The CODE project is founded by the European Union – Next Generation EU, Mission 4 "Education and Research" - Component C2 - Investment 1.1 - "Progetti PRIN 2022 PNRR".