Opening and Utilization of Public Data: User Review, Management, and Confidentiality from a Full Lifecycle Security Perspective

Authors: Gu Liping

Date: 2023-11-02T00:00:00+00:00

Abstract

This study investigates user qualification review, management frameworks, confidentiality agreements, and lifecycle security management for special data, emphasizing the critical importance of open data catalog review. The value and target selection for public data open utilization must comprehensively consider requirements from data processing and management policies, ensure data accuracy and reliability, and promote collaborative progress in scientific research and social development. Strengthening security management throughout the entire lifecycle of scientific data necessitates classifying data according to the varying confidentiality levels of scientific data generated at different stages of the research lifecycle, and establishing corresponding confidentiality review systems. The significance of cybersecurity assurance systems must also not be neglected, requiring enhanced technology research, development, and application to ensure the security and reliability of public data.

Full Text

Public Data Open Utilization: A Full Lifecycle Security Perspective on User Review, Management, and Confidentiality

Gu Liping[1][2]
[1] National Science Library, Chinese Academy of Sciences, Beijing, 100190
[2] Department of Information Resource Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing, 100190

Abstract: This paper examines user qualification review, management systems, confidentiality agreements, and lifecycle security management for special data, emphasizing the importance of data open directory review. The value and

target selection of public data open utilization must comprehensively consider data processing and management policy requirements to ensure data accuracy and reliability while promoting the common progress of scientific research and social development. Strengthening full lifecycle security management of scientific data requires classifying data according to different confidentiality levels generated at each stage of the research lifecycle and establishing corresponding confidentiality review systems. The importance of network security guarantee systems cannot be ignored; it is necessary to strengthen technology research, development, and application to ensure the security and reliability of public data.

Keywords: Public data; User review; Data management; Confidentiality agreement; Full lifecycle safety

---

1. User Qualification Review Methods for Special Data

User qualification and data security protection are important prerequisites for utilizing public data. Users must submit information regarding their purpose of use, qualifications, and confidentiality conditions, which undergoes strict review by the data center to ensure data security and privacy. Simultaneously, data centers must consider how to balance the relationship between open data utilization and privacy protection.

For special data, users primarily need to submit information about their intended purpose, qualifications, and confidentiality conditions, which is then reviewed by data center management personnel who strictly control the scope of access. The Scottish Longitudinal Study (SLS) serves as an excellent example of sharing complex and sensitive data. The SLS compiles data from routine administrative and statistical surveys, including census data, vital events data (births, marriages, deaths), NHS Central Register (migration into and out of Scotland), and NHS data (cancer registrations and hospital discharges), to examine migration patterns, inequalities, health, family reorganization, and other demographic, epidemiological, and socioeconomic issues. This data represents a valuable source of information for social decision-making. To protect privacy, a series of measures have been implemented: first, only a small group of researchers have access to personal data obtained through supplementary labor plans; second, datasets are anonymized, with no names or addresses of surveyed individuals retained in the database; third, actual data is stored on an isolated network with password protection, accessible only from secured locations; fourth, a steering committee is responsible for maintaining and adopting every research proposal reviewed by the Research Council, and will not authorize any research that could potentially identify individuals; and fifth, the data is not publicly available. Additionally, access to data is strictly controlled—if researchers wish to conduct remote analysis, the supplementary labor plan center can run statistical programs on their behalf.

The value and target selection of public data open utilization require consideration of multiple factors. First, public data should be used to promote public interest and social development. Second, open data utilization must follow certain rules and standards to ensure data accuracy and reliability. Finally, data utilization should fully consider the protection of personal privacy and commercial secrets to avoid data misuse and privacy violations.

Future data governance must reflect the speed and direction of data analysis technology development to adapt to new demands in data utilization and protection. Simultaneously, the governance process must fully consider the balance between public interest and risk resistance, developing appropriate policies and measures to promote the rational utilization and protection of public data.

---

2. Management Systems for Special Data and User Confidentiality Agreements

In international scientific research cooperation projects, the value and target selection of public data open utilization have become critical issues. Research and education institutions affiliated with different countries must follow their national laws, regulations, and data management policies when sharing scientific data to ensure compliance and security. For example, the data management plan of the Inter-university Consortium for Political and Social Research (ICPSR) emphasizes the importance of ethics and privacy, requiring that scientific data generated by research projects must come from informed consent and that personal information confidentiality must be protected.

In the ethics and privacy section of the ICPSR Data Management Plan, scientific data generated by research projects must come from informed consent, meaning that the project's informed consent statement must not prohibit data sharing within academia, and consent must be obtained specifically for this aspect. Additionally, information leakage risk management methods must be provided. Research projects must remove any direct identifiers from data before depositing it with ICPSR. Once deposited, the data undergoes a processing procedure to protect the confidentiality of personal information. These processing measures include: (1) rigorous review to assess leakage risk; (2) if necessary, modifying data to protect confidentiality; (3) restricting access to datasets where leakage risk remains high; and (4) negotiating with data producers to manage leakage risk. ICPSR designates certified qualified data management personnel as managers for data being processed in its leakage risk management process. This data is processed and managed using virtual desktop technology in a secure, non-networked environment.

To realize the value of public data open utilization, research and education institutions need to develop appropriate data management policies and measures. These policies should include provisions on the scope of data sharing, purpose and manner of data use, and data security and privacy protection. Simultane-

ously, research and education institutions need to establish effective data sharing platforms and mechanisms to promote the rational utilization and sharing of public data.

Regarding target selection, research and education institutions must determine data sharing objectives and priorities based on the actual conditions and needs of research projects. For projects involving major national interests and social public interests, the goal of data sharing should be to promote scientific research and social development. For projects involving commercial secrets and personal privacy, the goal of data sharing should be to protect privacy and commercial interests.

The value and target selection of public data open utilization require research and education institutions to fully consider the actual conditions and needs of research projects when formulating data management policies and measures, and to follow relevant laws, regulations, and policy requirements to promote the rational utilization and protection of public data.

---

3. Lifecycle Security Management and Data Open Directory Review

The extent of data processing is influenced not only by data purpose and value but also by multiple factors related to data collection and management. The value and target selection of public data open utilization must consider these factors to ensure data accuracy and reliability. Different research purposes may require different data processing methods and standards, necessitating customized data processing based on specific circumstances.

The extent of data processing is affected by data purpose, value, and multiple factors in data collection. For example, instrument-generated data is influenced not only by instrument capacity but also by how the data will be used: for parsing and annotation, for geophysical variables, mapping into spatiotemporal grid spaces, or for modeling. The processing methods for instrument-generated data are also influenced by the value generated from its use. If the contribution is theoretical, different research purposes have different definitions for handling outliers, scaling, missing values, and interpreting anomalies. If exploring the possibility of extraterrestrial life, the data processing approach is to preserve the original, comprehensive, and complete signals as much as possible. Therefore, strengthening protective management such as authentication and authorization for data downloads to prevent malicious use is an important task for various international-level data centers.

For instance, the data service policy of the Incorporated Research Institutions for Seismology (IRIS) states that it uses automated data management systems to process big data, which can move data from field sensors to user desktops. This system complies with international standards for data access, software, and I/O, enables metadata management and local data synchronization, and

the data center actively participates in outreach, training, and data exchange activities.

The value and target selection of public data open utilization must comprehensively consider multiple factors in data processing and management policy requirements to ensure data accuracy and reliability and promote the common progress of scientific research and social development. To protect the security and privacy of public data, strengthening protective management measures such as authentication and authorization for data downloads is crucial. Various international-level data centers need to establish sound data management policies and standards to ensure data compliance and security. Additionally, data centers must actively conduct outreach, training, and data exchange activities to promote the rational utilization and sharing of public data.

Regarding strengthening full lifecycle security management of scientific data, the value and target selection of public data open utilization must consider the different confidentiality levels of scientific data generated at each stage of the research lifecycle. To avoid situations where openable data is not opened or non-openable data is leaked, important international-level data centers must classify data levels and establish corresponding confidentiality review systems.

Simultaneously, different processing methods must be adopted for different data levels to ensure data accuracy and reliability. These processing methods not only affect data preservation and long-term archiving but are also closely related to how future generations will use this data. Therefore, data version control has become an important component of data archives in scientific data management.

The Earth Observing System Data and Information System (EOS DIS) of the National Aeronautics and Space Administration (NASA) defines data processing levels as follows: (1) Level 0: Reconstructed, unprocessed instrument data at full resolution, with instrument effects partially or completely removed; (2) Level 1A: Reconstructed, unprocessed instrument data at full resolution, with time information and auxiliary information for annotation; (3) Level 1B: Results of dividing Level 1A data into sensor units; (4) Level 2: Geophysical variables with the same resolution and location extracted from Level 1 data; (5) Level 3: Variables with completeness and consistency that can be mapped to a unified spatiotemporal grid; and (6) Level 4: Model outputs or results from low-level data analysis.

In pursuing their respective data values, changes in the scientific lifecycle lead to the generation of different data types. These data types are closely related to data flows in the scientific lifecycle, requiring the formulation of corresponding data management policies and standards based on data type and confidentiality level.

Data processing methods themselves affect how and to what extent data can be preserved and archived long-term, which in turn is closely related to how and to what extent future generations can use the data. If data is preserved in a form close to its raw state, then the processing algorithms and documentation

for transforming it into research-usable scientific data also need to be stored. Conversely, if data is stored in an advanced research form, then clues for tracing back to the original data sources also need to be preserved. Therefore, data version control is an important component of data archives in scientific data management.

In summary, the value and target selection of public data open utilization must comprehensively consider data management and confidentiality review requirements at each stage of the research lifecycle to ensure data security and reliability and promote the common progress of scientific research and social development.

---

4. Network Security Guarantee System

The value and target selection of public data open utilization must fully consider the importance of network security guarantee systems, strengthen technology research and development and application, ensure the security and reliability of public data, and promote the common progress of scientific research and social development.

Network security guarantee systems aim to adopt safe and reliable products and services, and improve management measures for data control, attribute management, identity recognition, behavior traceability, and blacklisting. At international scientific data centers such as IRIS, the Data Provider Agreement states that providers must report discovered problems to network operators through the IRIS DMC, provide statistics on data volumes distributed to network operators, and, if requested by IRIS, provide reports to secondary IP addresses or email addresses. Such practices serve as warnings and reminders in public notices, while in actual system security architecture they highlight data traceability and attack traceability technologies, primarily focusing on constructing security protection systems that prevent tampering, leakage, attacks, and viruses.

Open scientific information poses challenges for system and hardware engineers, requiring the development of secure methods for sharing confidential, sensitive, and special-grade data (to prevent accidents and avoid deliberate attacks). A data-intensive future may increase information security concerns. Sensitive data leakage is an inevitable gamble. Holders of personal data face not only management challenges but also the risk that even with encrypted formats, de-anonymization techniques can add more detailed data. There are already signs of lagging digital security, with less than one-third of digital information receiving minimal security protection and only half of data that should be protected being protected.

As scientific data continues to increase, open scientific information places higher demands on system and hardware engineers, requiring the development of secure methods for sharing confidential, sensitive, and special-grade data to prevent

accidents and avoid deliberate attacks. A data-intensive future may increase information security concerns, necessitating strengthened technology research and development to improve data security and reliability.

The value and target selection of public data open utilization must consider multiple aspects of information security and data management. Keeping source code and system architecture confidential is not a reliable method for ensuring information security, as the foundation of modern cryptography lies in the assumption that potential attackers know the internal workings. Conversely, open source code and system architecture allow attackers to analyze vulnerabilities but also enable more thorough system testing, thereby improving system security. This approach of "openness ultimately leading to better security" can also be applied to scientific data.

While ensuring data security, data integrity and provenance must also be considered. The Shibboleth single sign-on system developed by the UK Joint Information Systems Committee (JISC) eliminates the need for user credentials from content providers and allows institutions to control user access permissions, representing an important measure for protecting personal data security. Simultaneously, ensuring data integrity and provenance is also an important motivation for creating secure systems. Therefore, standard protocols and commercial community practices must be adopted to ensure data reusability and security.

Scientist codes of conduct also encourage individual responsibility and compliance with security laws in all countries and regions. Professional scientist codes of conduct include provisions encouraging individual responsibility. Although reports such as the UK's Universal Ethical Code for Scientists suggest individual consideration, all scientists should sign agreements with their employers to comply with security laws in all countries and regions. Scientists should sign agreements with their employers to ensure compliance with relevant regulations, thereby safeguarding the security and reliability of public data.

In summary, the value and target selection of public data open utilization must comprehensively consider multiple aspects of information security and data management, adopt open and standardized approaches, and strengthen individual responsibility and regulatory compliance to promote the common progress of scientific research and social development.

---

5. Emergency Management and Data Backup Mechanisms

The value and target selection of public data open utilization must consider not only data collection, storage, and processing but also emergency management and data backup measures. Disaster recovery backup mechanisms at data centers and the adoption of authoritative, standardized scientific research data management rules are important means to ensure the universality of research

data. These measures can ensure that effective response measures are taken promptly when data risks occur, safeguarding the security and reliability of public data.

Emergency management and data backup measures have two pathways: first, disaster recovery backup mechanisms at data centers, including emergency management systems and off-site backup of important scientific data; and second, the adoption of authoritative, standardized scientific research data management rules as an important means to ensure research data universality, enabling data restoration through other channels in case of loss according to international standards.

Regarding disaster recovery backup mechanisms at data centers, institutional design in international data center development strategies and corresponding cooperation from scientific data depositors are required. For example, IRIS's development strategy states that IRIS maintains, upgrades, and replaces necessary data archiving systems to ensure the capability and resilience to meet societal needs. Its Data Provider Agreement specifies that data providers must routinely transcribe data to new media at regular intervals to ensure data security and permanent availability through periodic data management, maintaining multiple copies of datasets to prevent loss or damage to any single dataset.

Furthermore, the adoption of authoritative, standardized scientific research data management rules is also an important means to ensure research data universality. These rules can ensure the standardization and normalization of research data, improving data reusability and shareability, and further promoting the open utilization and value realization of public data.

Information service institutions may appropriately draw upon and adopt these universal standards and guidelines established by professional standards organizations when publishing and utilizing research data. Examples include: (1) For author identification, ORCID provides unique author identifiers for users; (2) For data citation, DataCite was established in 2009 by the German Scientific and Technical Information Services, British Information Services, Australian National Data Service (ANDS), and other institutions to register research data and assign permanent identifiers, aiming to make research data usable as independent, applicable, and unique scientific objects; and (3) For research information standardization, organizations such as OpenAIRE, CASRAL, and EuroCRIS exist. OpenAIRE is a foundational support institution for the European Commission's open policy, providing standard guidelines covering all types of research data for researchers, while EuroCRIS provides an information tool for research activity management to researchers, research managers, information service institutions, publishers, and government agencies.

The value and target selection of public data open utilization must comprehensively consider multiple factors including data collection, storage, processing, emergency management, and data backup. By adopting authoritative, standardized scientific research data management rules and strengthening disaster

recovery backup mechanisms at data centers, the security and reliability of public data can be further improved, promoting the common progress of scientific research and social development.

Data center confidentiality and security measures are essential to ensure the security of public data during open utilization. As public data continues to increase, data confidentiality and security issues have become increasingly prominent, making it crucial to adopt a series of measures to ensure data security.

First, user qualification review for special data must be conducted to ensure that only users with appropriate qualifications and permissions can access and utilize the data. Simultaneously, sound management systems, confidentiality agreements, and lifecycle security management mechanisms for data must be established to guarantee data security at the institutional level.

Second, data open directory review must be strengthened to ensure that opened data does not involve sensitive information or privacy leakage. By establishing network security guarantee systems and strengthening emergency management and data backup mechanisms, data security and reliability can be further improved.

Additionally, attention must be paid to issues concerning data universality rights and interests. Adopting standardized scientific data management norms can prevent data loss and ensure data traceability and reproducibility, which helps protect the rights of data owners while promoting data sharing and circulation.

The value and security objective selection of public data open utilization are interrelated. While promoting public data open utilization, we must attach importance to data confidentiality and security issues and adopt a series of measures to ensure data security and reliability. This helps establish a healthy public data ecosystem, promotes data sharing and circulation, and drives social progress and development.

---

[01] The Scottish Longitudinal Study (SLS). Guides and Resources [EB/OL].[2023-04-07]https://sls.lscs.ac.uk/guides-resources/
[02] ICPSR. Data Management Plans[EB/OL].[2023-03-08]http://www.icpsr.umich.edu/files/datamanagement/
All.pdf
[03] Incorporated Research Institutions for Seismology (IRIS), Data Services. IRIS Data Services Policy Regarding Redistribution of IRIS Data Policy Version 2.0 [EB/OL].[2023-03-08].https://www.iris.edu/hq/files/programs/data_{services}/policies/Redistribution_{
[04] Incorporated Research Institutions for Seismology (IRIS), Data Services.Data Provider Agreement For contributors of data to the IRIS DMC[EB/OL].[2023-03-08].https://www.iris.edu/hq/files/programs/data_{services}/policies/Data_{{{Provi
V1.5.pdf
[05] Incorporated Research Institutions for Seismology (IRIS), Data Services. Data Services [EB/OL].[2023-08-16].https://www.iris.edu/hq/files/programs/data{services}/policies/Strategic_
[06] Incorporated Research Institutions for Seismology (IRIS), Data

Services.Data Provider Agreement For contributors of data to the IRIS DMC[EB/OL].[2023-09-22].https://www.iris.edu/hq/files/programs/data{services}/policies/Data_{{{Provide
V1.5.pdf
[07] BraseJ. DataCite -A global registration agency for research data [EB/OL].[2023-08-16] .http://ideas.repec.org./p/rsw/rswwps/rswwps149.html
[08] Principe P., Rettberg N.. OpenAIRE Guidelines: Supporting Interoperability for Literature Repositories, Data Archives and CRIS, Procedia Computer Science, Volume 33, 2014, Pages 92-94, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2014.06.015.

Note: Figure translations are in progress. See original paper for figures.