Policy Research and Analysis of Disciplinary Data Repositories: A Case Study in the Life Sciences

Authors: Sun Yinan, Gu Liping, Song Xiufang, Liu Jingjing, Jiang Xian, Gu Liping

Date: 2016-06-04T00:00:00+00:00

Abstract

[Purpose] To focus on policies for life sciences data repositories and provide recommendations for policy implementation. [Method] Through manual reading and screening, we investigated 38 data repositories in the life sciences domain with explicit policy statements, focusing primarily on policy statements regarding data submission, data management, and data use. [Result] Different stakeholder groups (data submitters, data managers, and data users) of disciplinary data repositories have distinct data rights and interests management specifications. [Limitation] Only 38 cases in the life sciences domain were investigated; there was no analysis of temporal changes in policy elements, and the discussion of policy implementation details is somewhat insufficient. [Conclusion] A sound disciplinary data repository policy system should include: data submission policies (content definition, format specifications, source requirements, attribution statements), data management declarations (data disclosure, data registration, disclaimers, data version management), and data use specifications (data access, recommended data citation, data licensing).

Full Text

Preamble

Total No. 265, 2015, Issue 12

Policy Research and Analysis of Disciplinary Data Repositories: A Case Study of the Life Sciences

Sun Yinan[1],[3], Gu Liping[1], Song Xiufang[1], Liu Jingjing[1],[3], Jiang Xian[2],[3]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Wuhan Library, Chinese Academy of Sciences, Wuhan 430071, China)
[3](University of Chinese Academy of Sciences, Beijing 100049, China)

This article is one of the research outcomes of the "CAS Literature Publishing Field Talent Support Program" (Project No: 1434).

Corresponding author: Gu Liping, ORCID: 0000-0002-2284-3856, E-mail: gulp@mail.las.ac.cn.

Abstract

[**Objective**] This study focuses on policies for life sciences data repositories to provide recommendations for policy implementation. [**Methods**] Through manual reading and screening, we investigated 38 life sciences data repositories with explicit policy statements, focusing on policies related to data submission, data management, and data use. [**Results**] The stakeholder groups of disciplinary data repositories (data submitters, data managers, and data users) each have different data rights management specifications. [**Limitations**] This study only examined 38 cases in the life sciences field, did not analyze temporal changes in policy elements, and lacked discussion of policy implementation details. [**Conclusions**] A sound disciplinary data repository policy system should include: data submission policies (content definition, format specification, source requirements, and attribution guidelines), data management statements (data disclosure, data registration, disclaimers, and data version management), and data usage specifications (data access, recommended data citation, and data licensing).

Keywords: Research data; Data repository; Policy research; Data management service; Life sciences

Classification: G353.1; G250.76; Q1

---

1. Disciplinary Data Repository Policies as Drivers of Data Sharing in Academic Fields

Scientific data constitutes an essential and indispensable component of scientific research [1]. Currently, major international research funding agencies [2], renowned research and educational institutions [3], and leading academic journals [4] have all established data policies that require research data to be stored and disseminated through data repositories. Furthermore, the recently developed data journals have integrated with data repositories [5] to jointly promote open sharing of research data.

Data repositories are primarily categorized into four types: institutional repositories, disciplinary repositories, multidisciplinary repositories, and

project-specific repositories [6]. Disciplinary data repositories, designed for specific academic fields, offer more systematic and professional services, enjoy broad audiences, and provide targeted support, making them highly favored by researchers across different domains. The research focus of disciplinary data repositories lies not in their technical construction, but rather in establishing a trust relationship between data providers, data managers, and data users by clarifying the rights and obligations governing data use.

Currently, policy research on disciplinary data repositories is driven by three main contexts: (1) librarians require a basis when introducing data management services (DMS) and recommending repositories to researchers, prompting many libraries to develop DMS [7-10]; (2) libraries or data centers need comprehensive policy systems when building repositories, with numerous domestic experts actively engaged in this practical work [11-14]; and (3) as repositories serve as primary infrastructure for data publishing and dissemination, China has developed considerable theoretical and practical knowledge regarding data open sharing [15-18]. Research on disciplinary data repository policies can advance progress in these three areas.

This study focuses on cases from life sciences data repositories, systematically examining the rights and obligations of stakeholders across seven sub-disciplines to summarize policy elements and provide a management framework.

2. Research Design

2.1 Research Approach

This study systematically examines and observes policies from life sciences data repositories through a framework addressing three core questions: (1) As data submitters, what rights do they hold and what obligations must they fulfill? (2) As repository managers (the repositories themselves), what rights do they hold and what obligations must they fulfill? (3) As data users, what rights do they hold and what usage norms must they follow?

2.2 Research Objects

According to re3data.org records (as of July 29, 2015), there were 1,305 data repositories, with 648 in the life sciences domain, of which 572 were openly accessible [19]. Referencing Scientific Data' s recommended list and classifications, we selected 38 life sciences repositories with explicit website policies as case studies [20], as shown in . These repositories span seven categories: nucleic acid sequence, protein sequence, molecular and supramolecular structure, neuroscience, omics, taxonomy and species diversity, and life-science community resources.

3. Data Submission Policies

Data submission policies form the essential foundation for any repository policy. After submitting data, researchers are primarily concerned with whether their data will be adequately protected and reasonably disseminated—this constitutes the cornerstone of trust between submitters and repository managers. Our investigation reveals four key components:

First, repositories must clearly define accepted content. Many life sciences subdisciplines provide detailed specifications. For instance, DGVa only accepts processed structural variation data along with study-related information including experimenter/samples, experimental protocols, and applied analyses; PDBe AutoDep accepts biomolecular NMR data; CXIDB accepts coherent X-ray diffraction imaging (CXI) experimental data; and GEO accepts gene expression and hybridization array data. Due to the specialized nature of life sciences, different repository categories store different data types, resulting in varied content specifications.

Second, repositories should specify required data formats based on community-standard software, tools, and transmission paradigms. For example, DGVa requires Excel spreadsheets and/or tab-delimited text files; UniProt encourages submissions in UniProtKB/Swiss-Prot format with textual descriptions; BMRB requires the NMR-STAR exchange format; and EuPathDB demands FASTA files for all available sequences, GFF files with complete gene information, and other format files.

Third, repositories must require submitters to comply with scientific ethics. Submission policies typically include statements prohibiting data from unethical experiments or collections, ensuring repositories do not disseminate products that violate scientific conscience. For example, GenBank stipulates that submitted human sequence data must not include any information that could reveal the donor' s identity.

Fourth, repositories need clear attribution guidelines. Copyright includes moral rights (inalienable, such as attribution) and property rights that repositories must explicitly address. Repository management teams should clarify how submitted data will be handled. For example, ArrayExpress specifies that submitted data are managed by specialized curation teams or systematically imported weekly from NCBI' s Gene Expression Omnibus.

4. Data Management Statements

Beyond data submission and management, repositories must develop usage specifications to protect submitters' rights and fulfill responsibilities to users, thereby fostering healthy data sharing in academic communities.

4.1 Encouraging Timely Data Disclosure

Data disclosure pathways and methods form the foundation of repository management policies, as all usage, dissemination, and modification behaviors depend on proper disclosure. Three primary pathways exist: (1) submitters receive account credentials for personal or group use; (2) besides submitters, academic peers can access data for peer review purposes; and (3) after submission, data become fully open for anyone to use. Repositories may employ one or multiple pathways, involving disclosure timelines and policy considerations. For example, EMDB encourages submitters to release data promptly rather than embargoing it.

4.2 Assigning Accession Identifiers

Just as academic papers are identified by author, title, journal, volume, and page numbers across platforms, research data require authentication mechanisms for persistent identification. Three common approaches exist: generic digital resource identifiers like DOI [59]; author identifiers like ORCID [60]; and repository-specific URNs or accession numbers. In life sciences, accession numbers have become a de facto standard for identifying research data, with many repositories establishing recognized allocation mechanisms. For instance, DGVa assigns unique, stable identifiers to study objects, variant regions, and sample-level variants after archiving; wwPDB and BMRB similarly allocate accession numbers. This practice has become standard for researchers when submitting journal articles, publishing data papers, annotating, and citing datasets.

4.3 Stating Disclaimers on Use Risks, Property Protection, and Terms

Internet content services typically include various statements. This study focuses on repository-specific disclaimers, which are particularly important in life sciences due to diverse data content, sources, and complex properties. Most surveyed repositories include relevant disclaimers covering three key aspects: (1) **Data use risk disclaimers**, where repositories state they make no warranties regarding software or data suitability and accuracy, and users assume all risks; (2) **Data intellectual property disclaimers**, where repositories do not evaluate the validity of patents, copyrights, or other IP rights on submitted data, nor guarantee that software or data use will not infringe third-party rights; and (3) **Data access process disclaimers**, where repositories assume no liability for user misuse or problems arising from database-related browsers, clients, or third-party software. Additionally, typical terms of use (e.g., prohibiting bulk downloads) and privacy terms (e.g., proper handling of usage logs) are included. Such disclaimers are essential management components, with use risk and IP disclaimers being particularly specialized and warranting further codification based on circumstances.

4.4 Requiring Submitters to Publish Updated Data Versions

Repository data typically falls into three scenarios: (1) static data from one-time experiments or observations that are final upon submission; (2) dynamically updated data that accumulates substantial content over time, requiring version identification (V1, V2, V3···Vn); and (3) data with varying completeness levels due to different publication stages or disclosure requirements, necessitating version differentiation and management. A fourth special case involves data requiring updates due to discovered errors, flaws, or omissions. Implementation requires attention to four issues: (1) repositories should present only the latest version while retaining all previous versions—INSDC states that while corrected data are removed from subsequent versions, all data remain accessible via accession numbers, with some repositories assigning new accession numbers to corrected files to ensure persistent access to both original and revised datasets; (2) each update or modification requires explanatory documentation—ClinicalTrials.gov mandates complete documentation when modifying or reassigning versions; (3) repository staff may modify files under special circumstances—BMRB allows staff to update citations or upgrade formats but requires revision statements at the file header; and (4) new versions should supplement rather than replace original datasets—PRIDE provides FTP details for adding data to existing datasets rather than requiring full resubmission. Disciplinary repositories typically require submitters to publish the latest version while avoiding overwriting existing datasets.

5. Data Usage Specifications

5.1 Clarifying Data Access Rights and Responsibilities

Repository users can typically upload, browse, download, and analyze data using provided tools, though some repositories differentiate user types (registered, advanced, and general users). While some require registration for uploading or downloading, others mandate institutional membership (with maintenance fees), and some allow open access without registration. Data access policies are common across nucleic acid sequence, protein sequence, molecular and supramolecular structure, neuroscience, omics, and life-science community resource repositories. Generally, repositories support open sharing and free download, use, or dissemination, with special cases subject to separate review processes. Some repositories also consider commercial contracts, funder requirements, or project protection needs that may restrict certain uses. Additionally, some allow programmatic development of analysis pipelines, websites, or data views for easier access and integration, such as EMBL-EBI. When users need access to protected datasets for verifying research results, repositories must protect provider rights, typically requiring formal applications. Finally, repositories must state that using data for patent applications requires provider permission or institutional/commercial use licenses, as specified in TCIA' s policies.

5.2 Standardizing Data Citation Formats

Data collection, curation, aggregation, analysis, and provision constitute vital scientific contributions that should be recognized when others use these data for new research. Data citation thus plays a crucial role in improving research evaluation systems and ensuring repository sustainability. Citation practices vary: some require dataset identifiers (digital resource identifiers or repository accession numbers), others require repository names and URLs, and some require citing a paper by the repository creators. Life sciences repositories generally prefer citing the repository and its information. For example: (1) ITIS format: "Retrieved [month, day, year], from the Taxonomic Information System online database, http://www.itis.gov" ; and (2) IRD recommendation: "Squires et al. (2012) Influenza Research Database: An integrated bioinformatics resource for influenza research and surveillance. Influenza and Other Respiratory Viruses DOI: 10.1111/j.1750-2659.2011.00331.x."

5.3 Specifying Data Licensing Agreements

Data licensing agreements constitute rights specifications for open sharing, safeguarding the scientific, economic, and social value of research data. These comprehensive specifications cover permitted activities: browsing, downloading, copying, redistributing, reusing, mining, extracting, compiling, deriving, sublicensing, and commercial use. Repositories primarily focus on licensing agreements for metadata and data, commonly employing Creative Commons licenses. Policies from life-science community resource repositories specify: (1) prohibiting profit-driven distribution (e.g., MGI requires explicit written permission for commercial use); (2) prohibiting commercial information extraction (e.g., ClinicalTrials.gov bans email address extraction for marketing); (3) special authorization for confidential information (e.g., NAHDAP requires researchers to protect personal confidentiality and prevent unauthorized use); and (4) non-exclusive rights (e.g., ClinicalTrials.gov states users own no proprietary rights and cannot represent the database).

6. Conclusion and Recommendations

(1) Policy Management Framework

Given varying academic norms across disciplines and even sub-disciplines within life sciences, policy research should focus not on "what policies should be" but on "what factors policy-making must consider." We recommend implementing measures based on life sciences research characteristics using the management framework in , which includes: content definition (specifying stored data types); format specification (following community standards); source requirements (complying with ethics); attribution guidelines (protecting moral and property rights); data disclosure (maximizing open sharing); data registration (allocating accession IDs); disclaimers (covering use risks, IP, access processes); version management (requiring documentation for updates); data access (gener-

ally free with special review processes); recommended citation (repository name, URL, and accession ID); and data licensing (typically CC/BY with emphasis on confidentiality and non-exclusivity).

(2) Practical Recommendations

This study offers two practical implications: institutions establishing disciplinary or institutional repositories for scientific data can reference our management framework (), and librarians promoting suitable repositories to researchers can evaluate options based on disciplinary classification and policy completeness.

(3) Research Limitations

Our conclusions require additional validation before broader application. When developing policies in China, we recommend incorporating researcher needs and consulting policy experts. Furthermore, this study did not analyze temporal changes in policy elements or examine implementation details, limitations that warrant future discussion with domain experts.

(4) Future Research

Building on this study, future work should combine researcher interviews to develop practical guidelines such as "Disciplinary Data Management Plan Guidelines."

Acknowledgments: The National Science Library, Chinese Academy of Sciences compiled the "Life Sciences Data Repository Case Studies" report, openly shared via the institutional repository (ir.las.ac.cn), which provided valuable reference for this research.

References

[8] Si Li, Xing Wenming. Scientific Data Management and Sharing Policies in Foreign Countries: Investigation and Inspiration to Us [J]. Information and Documentation Services, 2013, 34(1): 61-66.

[9] Borgman C L. The Conundrum of Sharing Research Data [J]. Translated by Qing Xiuling. New Technology of Library and Information Service, 2013(5): 1-20.

[10] Ku Liping. Basic Framework of the Analysis on Research Data Rights [J]. Documentation, Information & Knowledge, 2014(1): 34-51.

[1] Liu Rong, Zhang Na. Sharing Reflect the Value of Research Data—Visit to Chinese Academy of Engineering Sun Jiulin [J]. Science Technology Innovations and Brands, 2011(7): 10-13.

[3] Davidson J, Jones S, Molloy L, et al. Emerging Good Practice in Managing Research Data and Research Information within UK Universities [J]. Procedia Computer Science, 2014, 33: 175-182.

[4] Scientific Data to Complement and Promote Public Data Repositories [EB/OL]. [2015-09-07]. http://blogs.nature.com/scientificdata/2013/07/23/scientific-data-to-complement-and-promote-public-data-repositories/.

[5] Liu Fenghong, Cui Jinzhong, Han Fangqiao, et al. Data Paper: New Types of Academic Papers Published in Big Data [J]. Chinese Journal of Scientific and Technical Periodicals, 2014, 25(12): 1451-1456.

[6] Pampel H, Vierkant P, Scholze F, et al. Making Research Data Repositories Visible: The re3data.org Registry [J]. Translated by Ku Liping. New Technology of Library and Information Service, 2014(3): 26-34.

[7] Huang Ruhua, Qiu Chunyan. Research on Metadata Application Practices of Library Participating in Data Curation [J]. Library & Information, 2014(5): 65-69.

[11] Zhu Yunqiang, Sun Jiulin, Feng Min, et al. Study on e-Science for Geo-sciences [J]. Bulletin of Chinese Academy of Sciences, 2013, 28(4): 501-510.

[12] Ma Juncai, Liu Bin, Wu Linhuan, et al. Promote Microbiological Research and Application by Virtue of e-Science [J]. Bulletin of Chinese Academy of Sciences, 2013, 28(4): 519-524.

[13] Li Jianhui. Scientific Data Construction and Sharing of China [EB/OL]. [2015-07-20]. http://ir.las.ac.cn/handle/12502/7444.

[14] Liu Feng, Zhang Xiaolin, Kong Lihua. Research Review on the Research Data Repositories [J]. New Technology of Library and Information Service, 2014(2): 25-31.

[15] Zhu Yunqiang, Sun Jiulin, Wang Juanle, et al. Study on Earth Data Science and Data Sharing [J]. Land and Resources Informatization, 2015(1): 3-9.

[16] Liu Chuang, Wang Jinnian, Lv Tingting, et al. Standard for GMS_AdmBnd An Updated Moderate Scale Administrative Boundary GIS Database of Great Mekong Subregion [J]. Geomatics World, 2010, 8(1): 17-26, 42.

[17] Zhang Jilong, Yin Shenqin, Zhang Yong, et al. Social Scientific Data Sharing and Serving—An Example of Fudan University Social Scientific Data Platform [J]. Journal of Academic Libraries, 2015, 33(1): 74-79.

[18] Qu Baoqiang, Wu Jiaxi, Zhao Wei, et al. Analysis of Information Services of Local Scientific and Technical Literature Sharing Platform Based on Website Information [J]. Journal of the National Library of China, 2012, 21(1): 68-72.

[19] re3data [EB/OL]. [2015-07-29]. http://www.re3data.org.

[20] Scientific Data [EB/OL]. [2015-07-29]. http://www.nature.com/sdata/data-policies/repositories.

[21] GenBank [EB/OL]. [2015-07-20]. http://www.ncbi.nlm.nih.gov/genbank/.

[22] DGVa [EB/OL]. [2015-07-20]. https://www.ebi.ac.uk/dgva.

[23] EMBL-EBI [EB/OL]. [2015-07-20]. https://www.ebi.ac.uk/.

[24] ENA [EB/OL]. [2015-07-20]. https://www.ebi.ac.uk/ena.

[25] Gene Ontology [EB/OL]. [2015-07-20]. http://geneontology.org/.

[26] INSDC [EB/OL]. [2015-07-20]. http://www.insdc.org/.

[27] NCBI Sequence Read Archive [EB/OL]. [2015-07-20]. http://www.ncbi.nlm.nih.gov/Traces/sra/.

[28] Jackson Laboratory [EB/OL]. [2015-07-20]. http://www.jax.org/.

[29] PDBe AutoDep [EB/OL]. [2015-07-20]. http://www.ebi.ac.uk/pdbe/.

[30] UniProt [EB/OL]. [2015-07-20]. http://www.uniprot.org/.

[31] CXIDB [EB/OL]. [2015-07-20]. http://www.cxidb.org/.

[32] COD [EB/OL]. [2015-07-20]. http://www.crystallography.net/.

[33] BMRB [EB/OL]. [2015-07-20]. http://www.bmrb.wisc.edu/.

[34] ChEBI [EB/OL]. [2015-07-20]. https://www.ebi.ac.uk/chebi/init.do.

[35] EMDataBank [EB/OL]. [2015-07-20]. http://www.emdatabank.org/index.html.

[36] PCDDB [EB/OL]. [2015-07-20]. http://pcddb.cryst.bbk.ac.uk/home.php.

[37] wwPDB [EB/OL]. [2015-07-20]. http://www.wwpdb.org/.

[38] NeuroMorpho.Org [EB/OL]. [2015-07-20]. http://neuromorpho.org/neuroMorpho/index.jsp.

[39] NITRC [EB/OL]. [2015-07-20]. http://www.nitrc.org/.

[40] OpenfMRI [EB/OL]. [2015-07-20]. https://openfmri.org/.

[41] ArrayExpress [EB/OL]. [2015-07-20]. http://www.ebi.ac.uk/arrayexpress/.

[42] GEO [EB/OL]. [2015-07-20]. http://www.ncbi.nlm.nih.gov/geo/.

[43] dbGaP [EB/OL]. [2015-07-20]. http://www.ncbi.nlm.nih.gov/gap/.

[44] DIPTM [EB/OL]. [2015-07-20]. http://dip.doe-mbi.ucla.edu/dip/Guide.cgi.

[45] PeptideAtlas [EB/OL]. [2015-07-20]. http://www.peptideatlas.org/overview.php.

[46] EGA [EB/OL]. [2015-07-20]. https://www.ebi.ac.uk/ega/.

[47] PRIDE [EB/OL]. [2015-07-20]. https://www.ebi.ac.uk/pride/archive/.

[48] ITIS [EB/OL]. [2015-07-20]. http://www.itis.gov/.

[49] GBIF [EB/OL]. [2015-07-20]. http://www.gbif.org/.

[50] MGI [EB/OL]. [2015-07-20]. http://rgd.mcw.edu/.

[51] EuPathDB [EB/OL]. [2015-07-20]. http://eupathdb.org/eupathdb/.

[52] FlyBase [EB/OL]. [2015-07-20]. http://flybase.org/.

[53] Xenbase [EB/OL]. [2015-07-20]. http://www.xenbase.org/entry/.

[54] NAHDAP [EB/OL]. [2015-07-20]. http://www.icpsr.umich.edu/icpsrweb/NAHDAP/index.jsp.

[55] IRD [EB/OL]. [2015-07-20]. http://www.fludb.org/brc/home.spg?decorator=influenza.

[56] ClinicalTrials.gov [EB/OL]. [2015-07-20]. https://clinicaltrials.gov/.

[57] BioGRID [EB/OL]. [2015-07-20]. http://thebiogrid.org/.

[58] TCIA [EB/OL]. [2015-07-20]. http://www.cancerimagingarchive.net/.

[59] Mao Jun, Meng Liansheng, Zhen Xihui, et al. Establish a Framework of Digital Resource Unique Identifier System in China: Strategy and Economics [J]. New Technology of Library and Information Service, 2005(2): 1-4.

[60] Liu Runda, Wang Yunhong. An Overview of Open Researcher & Contributor ID [J]. Information Science, 2013, 31(11): 86-90.

Author Contributions

Sun Yinan: Primary manuscript writing, policy detail analysis;
Gu Liping: Research design, information source provision, policy element analysis, final manuscript revision;
Song Xiufang: Selection and research of investigation objects;
Liu Jingjing: Provision of Scientific Data repository list, case collection, participation in discussion and revision;
Jiang Xian: Compilation of major life sciences repository cases.

Received: August 3, 2015
Revised: September 10, 2015

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