Assessment Information

CoreTrustSeal Requirements 2020–2022

Repository:	Population Health Data Archive
Website:	https://www.ncmi.cn/
Certification Date:	9 July 2021

This repository is owned by: **National Population Health Data Center**

Population Health Data Archive

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, National repository system; including governmental

Reviewer Entry

Reviewer 1

Comments:
ACCEPT

Reviewer 2

Comments:
Accept

Brief Description of Repository

The Population Health Data Archive (PHDA) is an important data repository of the National Population Health Data Center (NPHDC). NPHDC [1] is one of the 20 national scientific data center approved by both Ministry of Science and Technology and Ministry of Finance, and it belongs to the science and technology resource sharing service platform under the national science and technology basic conditions platform [2]. The competent department of NPHDC is National Health Commission, with Chinese Academy of Medical Sciences being the supporting institution. NPHDC set up Resource Management Department, Engineering Technology Department, and Research Service Department, which are responsible for resource management, system construction and maintenance, information security assurance, standard specifications construction, and external service and publicity work of PHDA [3]. The medical data sharing team in the Institute of Medical Information, Chinese Academy of Medical Sciences is mainly responsible for the construction, management and operation and maintenance of PHDA. Based on the principle of “efficient, unified, systemic, integrated, safe, regulated, on-demand sharing”, PHDA supports multi-source (including national budget supported science and technology projects, medical and health departments and institutions such as Health Commission, Food and Drug Administration, and Chinese Center for Disease Control and Prevention, society and individuals), multi-disciplinarity (including biology, basic medicine, clinical medicine, public health, pharmacy, traditional Chinese medicine and pharmacy, population and reproduction, psychology) and multi-type (including questionnaire data, specimen data, clinical data, laboratory test results, genetic test data, medical imaging data, etc.) of scientific data’s registration, submission, review, quality control, storage, long-term preservation, certification, analysis, mining, evaluation and data sharing and other whole chain management functions, continually to integrate and dig the value of data, innovate data service models, and provide users at all levels of society with comprehensive, multi-level and personalized population health science data sharing services.

[1] https://www.ncmi.cn/phda/aboutUs.html
[2] Ministry of Science and Technology & Ministry of Finance: Notice on Publishing the Optimized and Adjusted Member Unit List of the National Scientific and Technological Resources Sharing Platform: http://www.gov.cn/xinwen/2019-06/11/content_5399105.htm. In the attachment, 20 National Science Data Centers are listed as the member units of the platform by the Ministry of Science and Technology & Ministry of Finance, and among them the National Population and Health Data Center comes the 14th on the list.
[3] Annex I: The organization structure of the Population Health Data Archive (PHDA)

Reviewer Entry

Reviewer 1

Comments:
ACCEPT

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

NPHDC established the comprehensive Population Health Data Archive (PHDA), facing multiple target groups such as government management departments, scientific research institutions, medical and health institutions, enterprises, journal publishing, and individuals, to carry out the collection, storage, integration, certification, processing, mining and sharing of scientific data in the field of population health. Designated communities include users of government departments, medical institutions, scientific research institutions, pharmaceutical companies, educational institutions, and the general public, etc. Currently, the domestic users of PHDA have covered 23 provinces, 5 autonomous regions, 4 municipalities, and 2 special administrative regions (with Hong Kong, Macao and Taiwan included); foreign users mainly come from 60 countries and regions including the United States, France, the United Kingdom, Russia, Japan, Netherlands, South Korea and Brazil.

Reviewer Entry

Reviewer 1

Comments:
ACCEPT

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

A. Content distributed as deposited, B. Basic curation – e.g. brief checking; addition of basic metadata or documentation

Reviewer Entry

Reviewer 1

Comments:
ACCEPT

Reviewer 2

Comments:
Accept

Comments

PHDA provides data metadata registration, data upload, data submission, and multi-level data review, and the PHDA website provides instructions on the collection procedure of data from different sources [1]. PHDA follows the custom metadata specifications, users are mainly required to fill in the registration items and provide a comprehensive description of the basic information, description information, source information, service information, and associated information of the dataset at the dataset registration stage. PHDA provides real-time online verification of metadata registration items, to help users to standardize the data descriptions, while generates data quality reports for each dataset and feeds back to the user. PHDA organizes experts in different fields such as biomedicine, basic medicine, clinical medicine, pharmacy, traditional Chinese medicine, and pharmacy and public health, to review and evaluate data quality of the submitted datasets in terms of integrity, consistency, availability, reliability, correctness, and give review results and recommendations. The dataset that has not passed the review shall be revised according to the review comments and submitted for review again after revision. PHDA also supports data curation for the approved data. The curator will optimize the descriptions of the dataset, including necessary subject classification adjustments, theme classification adjustments, keyword optimization, and dataset name specifications. PHDA will be in close communication with the datasets submitters during the curation process, in order to gain their identification.
[1] https://www.ncmi.cn/phda/submit.html This page is data collection page of PHDA, which provides users with the introduction of the collection procedure of project source data and other sources data.

Reviewer Entry

Reviewer 1

Comments:
ACCEPT

Reviewer 2

Comments:
Accept

Insource/Outsource Partners. If applicable, please list them.

The Institute of Medical Information, Chinese Academy of Medical Sciences is responsible for R&D and construction of PHDA, providing technical support for PHDA operation services [1].
[1] Annex II: Project assignment paper

Reviewer Entry

Reviewer 1

Comments:
ACCEPT

Reviewer 2

Comments:
Accept

Summary of Significant Changes Since Last Application (if applicable).

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

Other Relevant Information.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

ORGANIZATIONAL INFRASTRUCTURE

1. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

NPHDC is the only one national-level scientific data platform in the field of population health, which is responsible for the collection, processing, storage, mining, and sharing of scientific data, such as the major national science and technology

projects, the science and technology programs, and the major public welfare projects. And the overall development goal of NPHDC is to "use population and health scientific data to serve the Healthy China and benefit people's lives" [1]. NPHDC has been committed to becoming the scientific data collection center of population health field in China, to reaching the domestically leading, world-class, authoritative, and high-quality level.

According to the “Measures for the Management of Scientific Data” [2] and the “Regulations on the Management of Human Genetic Resources of the People's Republic of China” [3] promulgated by the General Office of the State Council of the People's Republic of China, NPHDC establishes the Population Health Data Archive (PHDA) to undertake the collection, review, processing, storage, management, certification and sharing of scientific data in the field of population health, and ensure the long-term preservation and continuous management, which specifically includes the following aspects:

(1) Provide multiple data collection mechanisms and channels for multiple types and multiple disciplines scientific data in population health field;

(2) Provide review for the collected population health scientific data, including data integrity, standardization, authenticity, confidentiality, reliability, and availability;

(3) Provide full life-cycle management for the collected population health scientific data, including data classification, labeling, sorting, and secure storage;

(4) Provide scientific data storage and long-term preservation services to ensure that the submitted scientific data will not be lost due to technological development, and be permanently preserved;

(5) Provide data sharing and use services at different levels and permissions facing the industry and society;

(6) Provide population health field scientific data certification and unique identifier distribution, to promote the scientific data citation and sharing;

(7) Provide data collection services for population health projects and issue collection certificates to assist the acceptance of science and technology projects at all levels.

More detail information about PHDA can be found on https://www.ncmi.cn/phda/support.html?type=aboutus.

[1] https://www.ncmi.cn/phda/aboutUs.html

[2] Measures for the Management of Scientific Data. Chinese version on http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm

[3] Regulations on the Management of Human Genetic Resources of the People's Republic of China. Chinese version on http://www.gov.cn/zhengce/content/2019-06/10/content_5398829.htm

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Data management and long term preservation is explicitly mentioned in PHDA mission statement.

2. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Submitters need to fill in the data permission information when registering and submitting data in PHDA, such as data open permissions, data sharing methods, data use permissions, data charging methods, etc [1]. Submitters can apply for a data protection period to protect their rights and set a specific dataset opening date. The data center will not release metadata and data publicly before the open date. In principle, the maximum data protection period is 3 years. For special reasons, the submitter can apply for longer data protection which requires submission of relevant certificates. In terms of data use permission, PHDA internally built multiple data and software use permission lists and detailed information descriptions, to help users better set up appropriate data use permissions and also allow users to fill in their own permissions. PHDA recommends the use of data in the CC0 license agreement or CC-BY 4.0 license agreement to allow more users to disseminate and use the data [2].

PHDA clearly stated the rights and obligations of the data submitter in User Help page [3]. The submitter reserves the right to submit data in accordance with relevant laws and regulations. For the submitted scientific data, the submitter has publishing right, authorship right, right of modification, right to protect the data integrity, right of use, etc. The submitter shall agree that the "National Population Health Science Data Center" share the following rights in the data copyright: the right to edit the data, the right to copy among different media, the right to network dissemination of the scope of the data disclosed in the registration agreement, the right of multilingual translations, the right to convert to different formats and rights of printing.

At the same time, the data center is equipped with teams of professional reviewers and experts for data review. If the data review fails, PHDA has rights to refuse the data release, or remove the infringing data content and pursue legal responsibility. PHDA Strictly abide by the "Measures for the Management of Scientific Data", the "Regulations on the Management of Human Genetic Resources of the People's Republic of China", the "Method for the Ethical Review of Biomedical Research Involving Humans", the "Regulations on Laboratory Animal Management" and other policies and regulations for data management, to restrict data access and permission. PHDA conducts hierarchical management

based on data sensitivity, and adopts different security level control measures for different sensitive levels of population health science data in the key links of the data management life cycle. The access data is divided into open data, controlled data and closed data. Users can access public data in the repository, but they shall get approval to access the controlled data. Data users must not infringe the rights of any individual or entity, and PHDA does not grant users the ownership or intellectual property rights of any content, data, or materials that may be accessed in this repository. For data with a clear statement of rights, the data user is responsible for obtaining permission from the responsible owner before using the relevant data. For high-level sharing of scientific data, PHDA provides remote virtual desktops, and users can perform remote data access and data analysis. If the data user violates the use of controlled or public data, PHDA will take appropriate actions.

With regards to the use of datasets, PHDA declared as follows: in order to respect intellectual property rights and protect the rights of data creators and service providers, data users are requested to clearly indicate the data source and data creator in the data-generated research outcomes (including publicly published papers, treatises, data products, and unpublished research reports or data products, etc.). For reprinted data (second or multiple releases), the source of the original data should also be indicated. Regardless of whether the results are published or produced in Chinese or English, the source must be indicated in the citation format specification. Without the written permission of the data center, no organization nor individual may modify or sell any part of the aforementioned data products, services, information, and materials in any way or for any reason. Anyone who infringes on the copyright and other intellectual property rights of the center will be held accountable in accordance with the law, and hereby solemnly declare [4].

[1] Annex III: A screen capture of the data permission information submission page in PHDA

[2] For example, the data service information section of the dataset details page (https://doi.org/10.12213/11.A001L.202009.46.V1.0) shows that the data usage license agreement is CC BY-SA 4.0.

[3] Annex IV: A screen capture of “Rights and obligations of the data submitter” in PHDA.

[4] Annex V: A screen capture of “Data Usage Statement” in PHDA.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

3. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
ACCEPT

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

The "Measures for the Management of Scientific Data" [1], the "Regulations on the Management of Human Genetic Resources of the People's Republic of China" [2], and the "Measures for the Management of National science and technology resource sharing service platform" [3] formulated by the Chinese government to promote scientific data management and sharing provide powerful guarantee for the sustainable development of National Population Health Data Center.

National Population Health Data Center is one of the 20 national scientific data center approved by both Ministry of Science and Technology and Ministry of Finance. It belongs to the Science and Technology Resource Sharing Service Platform under the National Science and Technology Infrastructure and it is the only national level scientific data platform in the population health field [4]. NPHDC built the Population Health Data Archive (PHDA), as the only custodian for data resources, provides population health field scientific data long-term preservation service and open sharing service towards the whole society. In order to ensure the continued access of PHDA, the National Population Health Science Data Center established appropriate organizational structures, including expert committees, user committees, ethics committees, resource management department, engineering technology department and research services department.

The competent department of NPHDC is National Health Commission, with Chinese Academy of Medical Sciences being the supporting institution. The Ministry of Science and Technology provides guidance and scientific research project support for the construction and management of the center, and government budget funds provide financial guarantee for the long-term operation of the data center [5]. In general, Chinese government will not cut the financial support for NPHDC. Even when this happens, Chinese Academy of Medical Sciences as the supporting institution will continue to support the basic maintenance of NPHDC. If all the funds are cut off, National Science and Technology Infrastructure will take over the repository and keep it function. In addition, our metadata directory can be obtained through National Technology Resource Sharing Service Platform [6].

[1] Measures for the Management of Scientific Data. Chinese version on http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm

[2] Regulations on the Management of Human Genetic Resources of the People's Republic of China, Chinese version on http://www.gov.cn/zhengce/content/2019-06/10/content_5398829.htm

[3] Measures for the Management of National science and technology resource sharing service platform. Chinese version on http://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2018/201802/t20180224_138207.html

[4] Ministry of Science and Technology & Ministry of Finance: Notice on Publishing the Optimized and Adjusted Member

Unit List of the National Scientific and Technological Resources Sharing Platform■http://www.gov.cn/xinwen/2019-06/11/content_5399105.htm. In the attachment, 20 National Science Data Centers are listed as the member units of the platform by the Ministry of Science and Technology & Ministry of Finance, and among them the National Population and Health Data Center comes the 14th on the list.
[5] Annex VI: The cover page of the newest funding grant
[6] National Science and Technology Resource Sharing Service Platform https://www.escience.org.cn/metadata/catalogue, Our metadata catalog is located in the website's Resource Catalogue column: All Category - Scientific Data - Population Health Science Data catalogue.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

4. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

PHDA is committed to the open sharing of scientific data, and provides sharing service to the public according to different data sharing levels. However, as PHDA shoulders the responsibility of data collection of large number of national scientific projects, in order to protect the legitimate rights and interests of projects' institutions, PHDA allows data submitters to set data protection period. During the protection period, PHDA only release catalogue of scientific data; if the data submitter agrees to release the submitted scientific data in advance, it will be made public from the date of its consent.

NPHDC has formulated the "Agreement for Resource Collection and Release of National Population Health Data Center"[1] and "Agreement for Resource Storage of National Population Health Data Center" [2] in accordance with the "Measures for the Management of Scientific Data"[3], "Measures for the Management of National science and technology resource sharing service platform"[4] and other laws and regulations, as well as biosafety management regulations such as "Regulations on the Management of Human Genetic Resources of the People's Republic of China"[5], "Ethics Review Measures for Biomedical Research Involving Humans"[6], "Regulations on Laboratory Animal Management"[7]. When data submitters register, submit and release data resources through PHDA, they should carefully read and strictly abide by these agreements. The storage agreement clearly states the rights and obligations of PHDA and depositor. The collection and release agreement stipulates that the data submitters shall ensure that they have all the copyrights of the submitted data or the authorization to upload the data, and promise that the information and relevant certification materials provided are completely true and valid. The item 6 on the collection and release agreement elaborates on the obligations and responsibilities of data submitters, including the deprivation processing of the data. If there is no processing, the necessary explanations must be made about the sensitive information such as personal identity in the data. The item 3 on the agreement also clearly states that the resources applied for submission and release should comply with the relevant regulations of the country and the population health industry information security management, and PHDA will not accept and release resources that contain unspecified identifiable personal information (with disclosure risk) or resources that involve illegal, repeated uploads, and substandard quality control.

The PHDA personnel that conduct data management and data review all received professional trainings in data security, human genetic resources management, data literacy, etc., and all obtained Good Clinical Practice (GCP) certification issued by the China Food and Drug Administration.

NPHDC establishes Expert Committee [8] and Ethics Committee [9] to ensure that the data collection, review, storage, and sharing of PHDA meet the law requirements and ethical standards. The Experts Committee is composed of experts and scholars with international reputation and profound academic attainments in the medical and health field from both home and abroad, and they jointly participate in the research and demonstration of major issues such as the center's top-level design and development strategy formulation, and put forward development suggestions. The Ethics Committee is composed of experts and scholars in the fields of population health, information security, ethics, sociology, etc. The ethics committee is responsible for reviewing the scientificity and ethical rationality of scientific data in the field of population health, fully grasping the data security demands, and urging the center's data security work meet relevant national regulations; supervising and reviewing the informed consent in projects and researches to ensure the privacy and security of research subjects; regularly conducting training on ethics knowledge, laws and regulations to promote standardized carry-out of scientific data management and sharing.

[1] Agreement for Resource Collection and Release of National Population Health Data Center.
https://www.ncmi.cn/ReleaseAgreement.html
[2] Agreement for Resource Storage of National Population Health Data Center.
https://www.ncmi.cn/ResourcestorageAgreement.html

[3] Measures for the Management of Scientific Data. Chinese version on http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm
[4] Measures for the Management of National science and technology resource sharing service platform. Chinese version on http://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2018/201802/t20180224_138207.html
[5] Regulations on the Management of Human Genetic Resources of the People's Republic of China, Chinese version on http://www.gov.cn/zhengce/content/2019-06/10/content_5398829.htm
[6] Ethics Review Measures for Biomedical Research Involving Humans. Chinese version on http://www.nhc.gov.cn/fzs/s3576/201808/14ee8ab2388440c4a44ecce0f24e064c.shtml
[7] Regulations on Laboratory Animal Management. Chinese version on http://www.gov.cn/gongbao/content/2017/content_5219148.htm
[8] See the list of expert committees:https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2805
[9] See the list of ethics committees:https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2804

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

5. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

1. The construction organization is an authoritative institution in China

The supporting organization of NPHDC is the Chinese Academy of Medical Sciences. The Chinese Academy of Medical Sciences is the only national-level medical science academic center and comprehensive medical science research institution in China, founded in 1956. The academy is strong in scientific researches. The medical researches cover all fields of medical science, including basic medicine, clinical medicine, preventive medicine, pharmacy, and medicine-related biology, physics, chemistry and other related disciplines.

In order to promote the integration, long-term preservation, and open sharing of scientific data in the field of population health, NPHDC established the Population Health Data Archive(PHDA). The specific construction work of PHDA is undertaken by the Institute of Medical Information of Chinese Academy of Medical Sciences. The Institute of Medical Information is the engineering technology center of the national population and health science data sharing service platform, and is also the national-level medical information research center, biomedical information resource center, the medical sub-center of the National Science and Technology Library and Documentation Center and the Health and Biomedical Information Cooperation Center of the World Health Organization, having been committed to the construction of domestic and foreign medical and health literature information resources and data resources for many years.

2. Owing a professional team and enough staff

NPHDC team has 38 full-time staff members. All the staff is from multiple background majors such as medical informatics, data science, computer science, biomedicine, health management, and information science, covering business areas of data management, data processing, data review, data analysis and mining, information release, sharing services, system management and operation and maintenance, technology research and development, product research and development, and others.

3. Having stable financial support

The construction, operation, and maintenance funds of NPHDC and its PHDA come from national government's financial investment. Approximately 15 million RMB of funding from the Ministry of Science and Technology and the Chinese Academy of Medical Sciences is obtained every year, to guarantee the continuous functioning of NPHDC and PHDA. The newest funding grant is also in place for 2019-2020 (including funding of 17 million RMB) [1].

[1] Annex VI: The cover page of the newest funding grant

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

6. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

NPHDC has an expert committee [1] composed of experts and scholars with international prestige and profound academic attainments in the field of population, hygiene and health from both at home and abroad. They jointly participate in the research and demonstration of major issues such as the center's top-level design and development strategy formulation, and put forward development suggestions. The expert committee is composed of more than 20 experts in fields such as medical informatics, information science, data management, data science, computer science, biomedicine, and health management. NPHDC regularly hold expert meetings on resource construction planning, archiving construction optimization, and service themes construction to engage experts in discussions and provision of guidance.
NPHDC regularly organizes exchanges with scientific research institutions, experts and scholars in the field, including the United States National Library of Medicine (NLM), the German National Library of Medicine (ZB MED), the Vrije Universiteit Amsterdam, and so on. The relevant staff of the data center also regularly participate in seminars and academic conferences in related fields at home and abroad, such as CODATA (Committee on Data), Biocuration (Biomedical Curation Conference), MedInfo (Medical Informatics Conference), ICSTI (International Council for Science and Technology Information) and AMIA (American Medical Informatics Association), to provide the staff with opportunities to learn and communicate with experts and scholars from at home and abroad■and at the same time promotes recent development achievements of NPHDC. Seminars are regularly held in NPHDC that allow staff to discuss with research scholars in the field about the problems and challenges faced by current research and development, exchange experiences, and demonstrate technologies.
NPHDC regularly conducts PHDA database systems introduction and functions training for users who submit data sourced from scientific and technological projects and users who submit data generated from other sources, and invites

users to exchange use experience and give feedback [2]. For graduate students, national population health scientific data management trainings are conducted, and the data literacy and data capacity of graduate students are investigated through questionnaires. PHDA sets the data administrator contact number and technical support contact number to receive user feedback on the repository and services at any time, and solve problems and capturing users' requirements.
[1] See the list of expert committees:
https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2805
[2] Function introduction and training of PHDA: https://www.ncmi.cn/special/phda/video_list.html

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

7. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

1. PHDA strictly reviews the identity of the data submitter, to guarantees the authenticity of the data.

PHDA strictly reviews the identity of the data submitter. Data submitter should use real name for registration, and fill in the

registration application form online [1], which requires identification number, cellphone number, email address, etc. After registration, users can use cellphone verification code or email link to activate the account, and after passing PHDA administrator's review, users can use the account to start the data collection process; if data submitter needs to make the institutional data collection or project data collection, it is required to choose the institution and upload the "Employment Certificate". When registering an institutional account [2], it is required to provide necessary information and conduct a strict review to ensure the authenticity of the identity of the submitter.

2. PHDA strictly implements the following strategies in the process of data registration, storage, management, release, and sharing, to ensure the integrity and authenticity of data and data metadata.

Data submitters register metadata and submit data through PHDA website [3, 4]. The system provides standardization and integrity real-time verification on metadata registration items according to the "National Population Health Data Center Archive Metadata Specification" [5], to ensure each registered dataset includes necessary meta-data information and data files, thus guaranteeing the standardization of data formats as well as the integrity and readability of data contents. When the metadata registration is completed, the system will show a completion prompt of the metadata fill-in.

PHDA formulates data backup scheme: the raw data submitted will be stored, and the data copy will be adopted for the data review, curation, release and external service. PHDA refers to relevant international and field standards (e.g. ISO 14721-2012 Open Archival Information System (OAIS) Reference Model) and performs integrity check during data transmission. In the process of data storage and long-term preservation, verifiable and/or auditable checks are performed to detect changes or losses in the data file, in the copy or between copies (such as checksum and recalculation, stability check, identification of lost files, etc.), and data consistency is ensured by regularly checking MD5 values of data stored locally and remotely. PHDA staff also regularly do the sample check on the storage medium, to make sure data backup can be used during disaster recovery.

PHDA supports the capture and recording of all change operations (such as creation, integrity check failure, deletion, modification, addition, saving, etc.) and audit/source information of the person or behavior performing the operation; it supports the management and monitoring across multiple storage availability levels (such as online, near-line, off-line), to achieve the whole trace and traceability of data information. PHDA will not change the data content without the authorization of the data submitter. PHDA supports two forms of data changes■

(1) Data update. Users can apply online, and upon approval they can modify the metadata and data files of the published dataset. After the modification, the unique identifier of the dataset remains unchanged. The repository records versions before and after the update and the description of the content change.

(2) Data version upgrade. It refers to editing greatly and modifying the existing dataset to a new version. PHDA assigns a new DOI and a new CSTR unique identifier for the new version of the dataset. The repository links up different versions of datasets, to facilitate data traceability.

[1] Personal account registration page: https://www.ncmi.cn/phda/register.html?active=user

[2] Institution account registration page:
https://www.ncmi.cn/phda/register.html?active=organization

[3] Data collection page: https://www.ncmi.cn/phda/submit.html

[4] Annex VII: The data collection procedure in the Population Health Data Archive

[5] National Population Health Data Center Archive Metadata Specification.
https://www.ncmi.cn/shareDocument/findContentManagementStandardDetail.do?id=361

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

8. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The PHDA dataset description adopts the "National Population Health Data Center Archive Metadata Specification" [1], which includes five parts: basic information elements set, description information elements set, source information elements set, service information elements set, and associated information elements set. This metadata standard is created by the NPHDC in combination with the characteristics of the data collected in the field of population health, It also refers to the "Dublin Core Metadata Specification", "Technology Platform Resource Core Metadata GB/T 30523-2014", DATs Metadata, DataCite Metadata, and other general and field data standards. The PHDA metadata standard considers versatility at the dataset level, at the same time, for the description of multi-source and multi-type datasets, it also considers the characteristics of medical datasets, including species classification, body system distribution, and subject terms and other elements descriptions.
To improve the discoverability and utility of data, and help data users discover and understand data from different perspectives, NPHDC provides users with high-quality data services, and configures multi-dimensional, fine-grained navigation and browsing, data retrieval, and data filtering functions based on registered metadata [2], including project

type, body systems, geographic mapping, project start and end time, theme words, etc.

PHDA provides web-based interactive collection services for data submitters. During the online data collection process, the repository provides annotations and sample explanations on the meaning of the registration items, helping data submitters to better conduct project metadata and data registration of metadata. The repository provides standardization and integrity real-time verification on metadata registration items to ensure each registered dataset includes necessary meta-data information and data files, while ensuring that the submitted metadata meets the established "National Population Health Data Center Archive Metadata Specification". PHDA provides a list of preferred file formats for different types of scientific data, which improves the standardization level of data storage, management and sharing [3]. Entity data files are recommended to be stored in parseable data formats such as XLS, CSV. If users submit data with specific proprietary format, PHDA recommends using the widely accepted general data format. For example, gene annotation files are usually stored in GFF, GTF, BAM and SAM formats; Sequencing files are usually stored in GBFF, FASTA and FASTQ formats. For this kind of data, users are required to provide corresponding file opening tools, or convert the file into general format, so as to facilitate data quality audit

PHDA formulates the "PHDA Data Resource Collection and Preservation Development Policy" [4] to guide the collection and archiving of data resources in PHDA. This policy includes contents of policy goal, collection scope, collection types, collection methods, collection languages, collection formats, collection standards, data services, policy updates, etc. If the data does not fall within the mission/collection profile of PHDA, the PHDA administrator will contact the data submitter and recommend other appropriate scientific data repositories.

For special reasons, the data submitter can apply for the cancellation of the submitted data, and NPHDC will assess whether it can be cancelled. For some special reasons, the data submitter may apply for revocation of the submitted data, and NPHDC will assess whether to revoke it. If an application for revocation is made for the submitted “project data”, the certificate of approval for revocation provided by the competent department shall be issued. In addition, once the scientific data resources registered in PHDA are found to be fraudulent and violate the requirements of relevant laws and regulations, the data submitter’s registration qualification will be revoked, and it will be notified on the website of PHDA system. What is more, the issued unique resource identification, certificate and other supporting documents will also be revoked. For the revoked dataset, its DOI will be directed to a tombstone page, which contains the basic description information for the revoked dataset [5].

[1] National Population Health Data Center Archive Metadata Specification. https://www.ncmi.cn/shareDocument/findContentManagementStandardDetail.do?id=361

[2] Data browsing page of PHDA: https://www.ncmi.cn/phda/browse.html

[3] Annex VIII: A screen capture of “preferred formats list” in the PHDA (https://www.ncmi.cn/phda/support.html?type=guide05).

[4] Annex XV: PHDA data resource collection and preservation development policy. This document describes the policy goal, collection scope, collection types, collection methods, collection languages, collection formats, collection standards, data services and policy updates of PHDA (https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2820) .

[5] DOIs tombstone page: https://www.ncmi.cn/dataSearch/getsearch_projectOrOtherHadData.html?type=1

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

9. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Based on the OAIS model, and combined with the characteristics of the population health scientific data management, PHDA established “PHDA Management Process” [1], which in detail described 6 functional entities of data ingestion, data archiving, data management, data access, system management and preservation planning, the technical strategies adopted in each process, and how staff manage these processes.
PHDA adopts distributed cluster architecture to deploy servers and storage devices, uses the multi-level storage strategy of online-nearline-offline. On-line storage adopts SAN storage and distributed storage, while near-line storage and off-line storage adopts tape libraries. PHDA formulates hierarchical storage strategy according to multidimensional factors such as source, content, and sensitivity of scientific data, to separate data storage from data application thus ensuring data security. After the data is transmitted to PHDA, the raw data of the dataset is stored, and data copies are used for data review, curation and external services.
According to the national standard of the People's Republic of China " Information security technology - Disaster Recovery Specifications for Information Systems (GB/T 20988-2007)", PHDA formulated data backup scheme [2] and deployed backup system to back up scientific data, software source code, system configuration parameters and log files, so that it can be restored in time once any problems occur in the programs or data, thus ensuring the continuous operation of basic

service when disaster happens. PHDA adopts a 5-level backup strategy, the first level is ordinary data, which adopts local backup; the second level data adopts basic remote backup; the third level data adopts regular network backup; the fourth level data adopts real-time remote backup; the fifth level data adopts real-time synchronization backup. Before data backup, the system automatically performs virus scanning and data verification to the files to ensure the stored backup data is accurate, and verifies data consistency by calculating the file checksum (MD5). The management personnel of PHDA regularly carry out sampling inspection of storage media to ensure backup availability during disaster recovery, including equipment patrol inspection once a week, and provide patrol inspection report and archive it after completion; preventive health check once a month, and health check report shall be provided and archived after completion [3].
[1] Annex IX: PHDA Management Process. This document describes the technical strategies adopted by PHDA in the short, medium, and long-term preservation of population health science data, and how NPHDC staff manage these processes (https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2817).
[2] Annex X: Data backup scheme. This document describes the data backup classification specifications, backup methods, backup strategies, backup tool requirements, etc (https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2818).
[3] Annex XI: Equipment inspection record. This document contains the storage and backup equipment inspection record, the monthly operation and maintenance inspection record, the hardware failure record, the server and virtualization inspection record and network security equipment inspection record.

Reviewer Entry

Reviewer 1

Comments:
Improved documentation, found a tool that translated much of the text in the PHDA management plan provided in appendix to English, that helped for several questions. Could have been 4.

Reviewer 2

Comments:
Accept

10. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Accept

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

PHDA supports the storage, management and long-term preservation of entity data, metadata and other related data description information. PHDA data resource storage and long-term preservation strategies are in Chapter 5 of “PHDA Management Process” [1]. PHDA determines the data sensitivity level (including non-sensitive data, low-sensitive data, moderately sensitive data, highly sensitive data, and extremely sensitive data, a total of 5 levels) according to the population health science data type, data volume, whether it involves sensitive information such as user privacy, and the degree of harm after the data is destroyed, to manage the grading and classification of data storage, and set different security levels of control measures (including time, space, personnel and technology) in each key link to ensure the data storage security in the PHDA.

Chapter 3 of "PHDA Management Process" stipulates the Submission Information Standard, the Archival Information Standard and the Dissemination Information Standard. The data submitter puts the data metadata and data files into Submission Information Package (SIP), and then submits the data through the PHDA system. The data submitter guarantees the reliability, authenticity, and validity of SIP, and promises to comply with the relevant national and population health industry information regulations on safety management. PHDA is responsible for virus detection and decompression of the received data, and then quality review is performed, including machine review and manual review, to ensure the quality of data acquired from various data sources. After the data inspection and preprocessing are completed, PHDA generates the metadata required by the system (including description metadata, technical metadata, preservation metadata, etc.) according to the Archival Information Package (AIP) structure defined by the system, and then classifies the received data objects and determines each object’s belonging set. Finally, PHDA uploads data to the preservation system for AIP archiving, updates the index, performs backups, and generates intake reports. To ensure data security, PHDA provides access functions based on the released database, which is a mirror image of the archive database. PHDA publishes dataset metadata, sample data, data dictionary, entity data, and related documents and tools within the scope of authorization according to user authorization and permission. The system provides complete, multi-dimensional data description, display, and browse, also provides version changes and traceability functions of data, data copyright and licensing, data citation and access methods.

Users who store data in PHDA must comply with the Agreement for Resource Storage of National Population Health Data Center [2], which imposes on the rights and obligations of the archive and the users who store data. The user side reserves the right to submit data in accordance with relevant laws and regulations. User side has the right to publish, authorize, modify, protect the integrity of the scientific data, and use the submitted scientific data. At the same time, the user side should agree that the "National Population and Health Data Center" also shares the right to edit, the right to copy among different media, the right of network dissemination in the scope of the data disclosure in accordance with the registration agreement, the right to multilingual translation, and the right to different formats conversion rights, and printing rights.

PHDA conducts regular data check, including the implementation of format identification, verification, and risk assessment

measures to ensure the sustainability of data formats. PHDA supports standardized format conversion of data resources in risk format, and migration of information to a new sustainable format. This process will be completed together with the data submitter. The relevant regulations of data format management are described in detail in section 9.1 of the "PHDA Management Process".

[1] Annex IX: PHDA Management Process. This document describes the technical strategies adopted by PHDA in the short, medium and long-term preservation of population health science data, and how NPHDC staff manage these processes (https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2817).

[2] Agreement for Resource Storage of National Population Health Data Center. https://www.ncmi.cn/ResourcestorageAgreement.html

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

11. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

PHDA uses the "National Population Health Data Center Archive Metadata Specification" [1] to describe the collected data, which mainly includes five element sets: basic information (such as registered dataset name, data identification, data type, data size, etc.), description information (such as keywords, abstract, subject classification, etc.), source information (such as creator, creating organization, contact person, contact information, etc.), service information (such as sharing method, sharing permission, and data permission, etc.), and associated information(such as related papers, supporting tools, etc.). The establishment of the metadata specification refers to the "Dublin Core Metadata Specification", "Technology Platform Resource Core Metadata GB/T 30523-2014", DATs Metadata, DataCite Metadata and other general and domain data standards. PHDA provides users with a friendly and interactive web interface, helps users in data and metadata registration and collection in standard formats, and guarantees the integrity and understandability of data and metadata to the greatest extent. In the metadata registration process, it provides grammar and form check for content filled in online, and provides real-time display of metadata integrity to show users the progress and completeness of the data.

PHDA adopts a combination of manual review and computer aided verification, and establishes multiple data quality control and review mechanisms at all stages of the data life cycle to ensure data quality, including data integrity, consistency, accuracy, reasonability, etc. In the data collection stage, PHDA requires data submitters to commit to the quality of the submitted metadata and datasets. In multiple key links such as project information registration, data metadata registration, data upload, data submission, and data service, PHDA implements data quality testing, control, and user interaction feedback strategies to form a closed loop of iterative feedback to optimize data filling, testing and submission, in order to effectively guarantee data quality [2, 3]. Quality review on the data from project source mainly includes five parts: project information review, project data metadata review, submitted data review, expert review and final review. While quality review on the data from other sources includes four parts: data metadata review, submitted data review, expert review and final review. Among them, project information review, project data metadata review and submitted data review are mainly completed automatically by the online data collection system; expert review is done by experts from different disciplines; final review is completed by data center administrators. PHDA provides process navigation and information prompts for users in each key link of registration and submission. The progress of data collection and related information feedback can be checked in the personal management center of registered users.

In order to promote data citation and sharing, PHDA standardized the data referencing format [4]. The referencing format of data citation is: Author or Institution, Resource name, Population Health Data Archive of National Population Health Data Center, Time, Unique Resource Identifier. This referencing format has been displayed on the details page of each dataset. In addition, PHDA also provides three citation formats to download, including EndNote XML, RIS and BibTex.

PHDA supports all-round and dynamic supervision and evaluation of stored scientific data and data resource services, promotes the continuous and sound development of data collection, and assists in the selection and development of high-quality resources. In order to ensure the authoritative and scientific nature of the data resources stored in PHDA, NPHDC established the "National scientific data sharing platform for population and health - Resource evaluation system" based on the perspective of resource management and following the principles of combining quantitative indicators with expert evaluation and the principles consistent with the evaluation system of the Ministry of Science and Technology. According to this resource evaluation system, PHDA organizes experts in various fields such as medical informatics, information science, biomedicine, and health management, from multi-dimensions like resource theme representativeness, authoritativeness and industry influence, resource quality, data scale, and service capabilities, to conduct dataset testing and evaluation, form institutional testing and evaluation results and analysis reports, and finally

release results to the public in a timely manner. In addition, for the published and shared data in PHDA, PHDA calculates the data score (U index) based on the data downloads, data clicks, the number of entity data applications, the number of data collections, the number of data citations, the number of data subscriptions, and the number of data sharing. The score is displayed as a label under the title of each dataset.

[1] National Population Health Data Center Archive Metadata Specification. https://www.ncmi.cn/shareDocument/findContentManagementStandardDetail.do?id=361

[2] Annex IX: PHDA Management Process. This document describes the technical strategies adopted by PHDA in the short, medium and long-term preservation of population health science data, and how NPHDC staff manage these processes (https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2817).

[3] Annex VII: The data collection procedure in the Population Health Data Archive (https://www.ncmi.cn/phda/submit.html)

[4] Annex XIV: A screen capture of “Reference format” in PHDA.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

12. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

PHDA established full life cycle data management process, covering data registration, data collection, data saving, data review, data integration, data management, and data access. PHDA carries out multi-ways for data gathering and collection of project source data and other sources data, establishes comprehensive data resource audit, quality control, and management system, and provides a variety of data authentication methods such as data resource unique identifier, data certificate, and project data collection proof. PHDA website clearly records all data processing workflows, and measures such as data security and privacy protection are taken throughout the workflow. PHDA sets up an administrator position, responsible for communicating with data submitters in the entire data management process, and organizing experts to conduct data review, and timely feedback results to the data submitters until the data is approved.

1. Data Collection

Detail descriptions of data collection workflow are shown in the PHDA data collection page [1]. PHDA strictly follows the workflow and conduct data collection, review, archiving, storage, and release of project source data and other sources data. Data submitter registers and logs in to the account, performs metadata registration, and uploads entity or sample data files, data dictionaries, and other related materials. PHDA provides real-time verification of metadata items and conducts quality review of data files, including data integrity, standardization, consistency, etc., to ensure that each registered dataset contains necessary metadata information and data files. For data that fails the check and quality review, PHDA will provide modification tips, and the user will modify the problem based on the feedback. After passing the system quality review, data submitter can submit the data, and the administrator will assign field experts according to the type of scientific data for expert review. Datasets approved by experts can be released in accordance with the data sharing agreement.

2. Data Quality Control and Review

PHDA adopts a combination of manual review, information system logic design, and auxiliary verification software, and establishes multiple data quality control and review mechanisms at all stages of the data life cycle, including but not limited to data integrity, consistency, accuracy, reasonability, completeness, etc. See R8 and R11 for detailed information.

3. Data Storage

PHDA supports long-term preservation of entity data, metadata, and other related data description information. PHDA supports comprehensive data backup and disaster recovery, including automatic backup of scientific data, software source code, system configuration parameters, and log files, to ensure long-term preservation of scientific data and continuous operation of sharing services. See R9 and R10 for details of data storage management.

4. Data Access

PHDA releases data in accordance with the data sharing agreement. Users can access the open data in the repository. For controlled data, it is needed to send an access request to the data provider, and it can only be accessed after approval. For scientific data of higher sharing level, PHDA provides remote virtual desktops, and users can perform remote data access and data analysis. When accessing, downloading, and using PHDA data, users must follow the corresponding data license and use the standard data referencing format specified by PHDA.

PHDA also provides users with high-quality data services. The system configures multi-dimensional, fine-grained navigation browsing and data filtering functions for registered metadata; supports retrieval functions for datasets, datasheets to data records, to satisfy users' one-stop retrieval requirements; supports automatic statistics calculation on data browsing, downloads, retrieval, applications, etc., which users can check in page "Personal Center - Statistical Information".

5. Data Security Control

The PHDA website clearly records all data processing workflow, and the data security and privacy protection measures are adopted throughout the workflow. It combines the type and volume of population health scientific data, whether the data is classified data, whether sensitive information such as user privacy is involved, and the degree of harm after the data is destroyed, to determine the sensitivity level of the data (including non-sensitive data, low-sensitive data, moderately sensitive data, highly sensitive data, and extremely sensitive data, a total of 5 levels), and carry out data security level tagging and implement 5-level security control measures. At present, the sensitivity level of data is judged by both machine and manually. In the process of data submission, machine evaluation of sensitive information identification of dataset metadata, data dictionary, and data file content is performed, and the evaluation results assist manual labeling of the sensitivity level of the dataset. PHDA clearly sets the security control mechanism (involving time, space, personnel and technology) of data with different sensitivity levels in each key link of the data management life cycle, including data collection, data transmission, data check, data storage, data processing, data access, data use and data destruction, etc., to ensure the security of data in PHDA.

6. Adjustment of data workflow

With the continuous development of science and technology, PHDA will inevitably adjust and update its workflow. The new workflow must be fully tested and modified before it is approved for official operation. The existing system and workflow will continue to be maintained until the test is completed. The PHDA website will simultaneously announce the change notice of the workflow and provide relevant guidance for users.

[1] Annex VII: The data collection procedure in the Population Health Data Archive (https://www.ncmi.cn/phda/submit.html).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

13. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

PHDA maintains a searchable metadata catalogue, which follows the FAIR scientific data management guideline——Findable, Accessible, Interoperable and Reusable. Through this metadata, PHDA supports multi-dimensional data feature disclosure, source tracing, version changes and service information. The PHDA system provides metadata downloads of datasets in XML format, and the download format follows the national standard "Technology Platform Resource Core Metadata GB/T 30523-2014 ". At present, PHDA is building mappings between other international and field metadata standards and PHDA metadata standards, to provide more ways of data acquisition. PHDA creates an API for machine-harvesting open dataset metadata. The metadata standard provided by the dataset complies with the national standard "Technology Platform Resource Core Metadata GB/T 30523-2014", which made it possible to perform the metadata collection to China Science and Technology Resource Sharing Network through API. Currently, PHDA is also building APIs that follow more metadata standards.

Based on the "National Population Health Data Center Archive Metadata Specification", PHDA provides users with multiple data discovery services, including primary and advanced retrieval of storage content collection. Primary retrieval includes retrieval of resources such as scientific data, shared documents, work news. Advanced retrieval [1] supports the search of the data form project source and data from other sources. Advanced search includes dataset name, dataset description, keywords, dataset identifier, datasheet name, creator, creator's name, creator ORICD, etc., providing fuzzy search, precise search and Boolean logic search. PHDA also provides data certificate query to help users perform dataset authentication and verification [2]. The data retrieval and discovery mechanism of PHDA runs through data sets, data tables and data records. While meeting users' one-stop data retrieval needs, PHDA also provides data filtering on species, data source, data level, sharing methods, data size, data format, etc., to strengthen data discoverability and utilization.

NPHDC automatically assigns dual identifiers to scientific data sets that passed the review, including Digital Object Identifier (DOI) and scientific and technological resource unique identifiers. The identifier does not change with the attributes of the identified object (e.g. copyright owner, storage address), which helps in the persistent identification and access of datasets. Users can accurately obtain the detailed information of dataset through DOI link, which is convenient for tracking the citation of scientific data, and promotes the interoperability of scientific data resources [3].

In order to promote data citation and sharing, PHDA standardized the data referencing format [4]. The referencing format of data citation is: Author or Institution, Resource name, Population Health Data Archive of National Population Health Data Center, Time, Unique Resource Identifier. This referencing format has been displayed on the details page of each dataset. In addition, PHDA also provides three citation formats to download, including EndNote XML, RIS and BibTex.

PHDA is a generalist repository in the field of population health, which collects and integrates data resources based on datasets. It supports the registration, collection, review, quality control, preservation, certification, analysis, mining, evaluation and sharing of multidisciplinary scientific data such as biology, basic medicine, clinical medicine, public health,

pharmacy, Chinese medicine, population and reproduction, and psychology. To promote open access to scientific data, PHDA has been registered in scientific research data knowledge base registration directory system “re3data.org” (Registry of Research Data Repositories) and bioinformatics knowledge base “FAIRsharing”. PHDA can be looked up in their website: re3data (https://www.re3data.org/repository/r3d100013199), FAIRsharing (https://fairsharing.org/biodbcore-001377/).

[1] https://www.ncmi.cn/phda/data/api.html PHDA provides API, which supports third parties to use this interface to integrate scientific dataset metadata from other source data into retrieval services on other service platforms.

[2] Annex XVI: A screen capture of “Advanced search example”.

[3] https://www.ncmi.cn/dataSearch/certificate_search.html

[4] For example, through the link https://doi.org/10.12213/11.A001L.202009.46.V1.0 can access the detailed information of the dataset "NSD3 directly binds to IRF3 and methylates K366 of IRF3".

[5] Annex XIV: A screen capture of “Reference format” in PHDA.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

14. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The construction of PHDA follows the FAIR Guiding Principles for scientific data management and stewardship, to keep data, metadata and other related data description information findable, accessible, interoperable and reusable.
The PHDA dataset description adopts the "National Population Health Data Center Archive Metadata Specification" [1]. When submitting data, data submitters shall comply with the standard and provide complete and authentic metadata description information. The standard includes five parts: basic information elements set, description information elements set, source information elements set, service information elements set, and associated information elements set. This metadata standard is created based on the characteristics of the data collected in the field of population health, with reference to the "Dublin Core Metadata Specification" and "Technology Platform Resource Core Metadata GB/T 30523-2014", DATs Metadata, DataCite Metadata, etc. Currently, PHDA is building mappings between "National Population Health Data Center Archive Metadata Specification " and these metadata standard core elements.
PHDA provides data users with comprehensive data metadata, sample data and data dictionary (describe the name, type and meaning of each data variable) to help users understand the meaning of the data and basic forms of the data. In order to help users better understanding and using data, PHDA performs descriptive statistical analysis on open entity data, and statistically describes related data of all variables of the canonical data tables.
In order to make scientific data easily accessed and displayed, and improve the standardization of data storage, management and sharing, PHDA will recommend compatible general formats to scientific data submitters, including the following types:
(1) For tabular data, PHDA recommends using parseable data formats such as XLS and CSV, and PHDA also provides sample data templates of formats to data submitters.
(2) For data with specific proprietary format, PHDA recommends using the widely accepted general data format. For example, gene annotation files are usually stored in GFF, GTF, BAM and SAM formats; sequencing files are usually stored in GBFF, FASTA and FASTQ formats; molecular biology data files are usually stored in PDB format; physiological information number file is usually stored in EDF format. In view of these data formats, PHDA requires that supporting software tools also be uploaded when submitting data to better support data analysis.
(3) For data without a specific general format, PHDA recommends using data formats that are easy to understand, access and use, such as general data formats for text in TXT, PDF, DOC and PPT; for images in JPG, PNG, TIFF; for audio and video in MPEG-3, AVI, MOV, etc., in order to avoid the risks caused by data storage format changes.
(4) PHDA has also developed a data conversion tool from RDB to RDF based on R2RML, which supports the conversion from relational database to semantic data, and PHDA is now preparing to integrate it into the repository.
Until now, PHDA never happened situations that need to upgrade or convert format of data or files due to invalid original format. If the storage format changes, the data in the new format will be used as new version of the dataset, and history versions of the data will still be permanently retained in the repository, at the same time association of different versions of the dataset is supported.
[1] National Population Health Data Center Archive Metadata Specification.
https://www.ncmi.cn/shareDocument/findContentManagementStandardDetail.do?id=361

Reviewer Entry

Reviewer 1

Comments:

Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

15. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The construction of PHDA and related hardware and software infrastructures refers to the international standards "ISO 14721 Space Data and Information Transfer Systems - Open Archival Information System - Reference Model" and "ISO 16363 Space Data and Information Transfer Systems - Audit and Certification of Trustworthy Digital Repositories", and the national standards "GB/T 22239-2019 Information Security Technology - Baseline for Classified Protection of Cybersecurity” and “GB 50174-2017 Code for Design of Data Centers”. As per the requirements of infrastructure part in the ISO 16363 standard, manage the number and location of all digital objects backups, and maintain systematic analysis for security risks factors that relate to data, system, personnel and hardware devices. According to the GB 50174-2017 standard and level 3 security requirements in GB/T 22239-2019 standard, build safe and reliable software and hardware infrastructures, such as computer room site, communications network, computing environment and archiving environment,

to guarantee the availability and stability of services. The PHDA systems will conduct self-inspection and third-party audit regularly, at least once every six months.

PHDA formulates an infrastructure development plan, continuously strengthens the construction of cloud platform, high-performance computing clusters, artificial intelligence GPU clusters, distributed storage system, servers and other infrastructures, provides software and hardware technologies that meet users' service needs, and continuously improves the computing, storage, and network capabilities to ensure maximum service availability and reliability. The existing more than 200 rack server, more than 500 virtual servers, 8 minicomputers, 8 big data all-in-one machines, and 6 GPU artificial intelligence servers ensure the computing power of PHDA. The existing storage system has a capacity of 1.5PB and the virtual tape library has a capacity of 600TB. The data is classified and stored according to multidimensional elements such as data source, content, format, value and sensitivity. Meanwhile, a distributed file system is configured to support Hadoop/MPP distributed computing and big data analysis and processing. The 10-Gigabit twin-engine core routing switch was used to establish a large-capacity and high-performance network environment with a 10 Gigabit Ethernet in the backbone area and Gigabit connection to the user end. The internet outlet independent bandwidth reaches 570 Mbps. At present, the total amount of data collection has reached 113TB, and the number of system visits has reached 890,000. The availability, bandwidth and connectivity of the system can meet the needs of users.

PHDA is developed based on B/S mode and J2EE framework, using Windows/Linus as underlying operating systems. Relevant data collection, storage, and management software systems are in use, such as MySQL database, MPP database, tape library management system, data backup system, vSphere server virtualization management system, Zabbix monitoring system, etc. PHDA provides API, which supports third parties to use this interface to integrate scientific dataset metadata from other source data into retrieval services on other service platforms. In addition, PHDA has professional operations and maintenance personnel responsible for the daily inspection, monitoring and maintenance of hardware and software equipment [1, 2], making every reasonable effort to maintain the continuity of online services and provide early warning of any changes or discontinuity.

To ensure continuity and accessibility of PHDA, the disaster plan and business continuity plan are devised, including building emergency management [3] and disaster recovery backup mechanism [4]. The emergency management system classifies and grades security events such as destructive attacks in the software system, database failures, equipment security failures, and network failures. The backup system automatically makes copies for scientific data, software source code, system configuration parameter and log files, so that it can be restored timely once any problems occur in the programs or the data. PHDA stores two copies of online data in real time on two sets of storage devices, realizing active-active data level operation and seamless data switching in case of hardware and software failure; through the high-speed optical fibre internet to achieve remote disaster recovery backup of data. Adopting ways of remote backup, heterogeneous backup and multiple backup can guarantee the continuous functioning of the basic service when any disaster happens.

[1] Annex XI: Equipment inspection record. This document contains the storage and backup equipment inspection record, the monthly operation and maintenance inspection record, the hardware failure record, the server and virtualization inspection record and network security equipment inspection record.

[2] Annex XVII: Network security situation monitoring record, including the monitoring of host situation and attack intrusion.

[3] Annex XIII: Emergency plan. This document describes measures to be taken in the case of storage equipment failure, network equipment failure or server failure, natural disasters (flood, fire, electricity, etc.) and other emergencies (https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2819).

[4] Annex X: Data backup scheme. This document describes the data backup classification specifications, backup methods, backup strategies, backup tool requirements, etc (https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2818).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

16. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
Accept

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

PHDA sets the information security management post, including many roles such as network administrator, system administrator, application administrator, security administrator, security auditor, computer room administrator, which is responsible for executing safety management and maintenance work of network, host system (including database), application system and computer room (including air conditioner and UPS). PHDA can monitor 7*24 on the security status of the network and equipment by the security platform, including attack intrusion, malicious code, host situation, website situation, asset situation, etc [1].
In PHDA, operations of data registration, upload, review, curation, storage, management, sharing and use leave traces

throughout the process, which is logging in the backstage system, including information of the operator, operation (including addition, deletion and modification), request parameters, etc. PHDA establishes security risk management system, standardizes security monitoring and auditing mechanisms, and ensures the stable and secure operation of the repository through risk identification, risk analysis, risk response and risk monitoring related technologies [2]. PHDA uses Zabbix to establish a business monitoring system to realize host performance monitoring, network device performance monitoring, database performance monitoring, FTP and other general protocol monitoring and provide failure warnings. In addition, PHDA establishes emergency management mechanism, constructs emergency management system, and performs categorization and classification of security incidents such as software system destructive attack, database failure, equipment security failure and network failure [3].

PHDA is constructed in accordance with the level 3 security standards in “GB/T 22239-2019 Information Security Technology - Baseline for Classified Protection of Cybersecurity” [4], divides network security domains, builds a security protection system with boundary protection, antivirus, anti-attack, anti-tampering and anti-leakage protection. In addition, PHDA implements five major types of technical measures of secure physical environment, secure communication network, secure area boundaries, secure computing environment, and security management center, and adopts management measures such as data control, attribute management, identity recognition, behavior trace, log analysis, security audits and blacklists, to conduct defense in-depth from the outside to the inside. PHDA also implements strict access control, with centralized management of operation and maintenance through VPN+ bastion host, and the adoption of two-factor authentication, combined with database audit, log audit and other equipment for security audit, to ensure network security, host security and application security. PHDA implements defense against various malicious attacks such as DoS/DDoS attacks, ARP fraudulent attacks, address sweep attacks, and port scan attacks through network firewalls; tracks and manages the number of website connections, website number, and website concurrency through WEB application protection; and regularly scan the system for security vulnerabilities through the host vulnerability scanning equipment and WEB application vulnerability scanning equipment.

PHDA adopts protocol OAuth2.0 to conduct the user identification authentication service. Through this protocol, third-party applications can access the protected resources stored by users on an application system in a secure way. The user authentication system collects as little as possible of the user's personal data and sets personal data expiration time. All the data transfer between application programs in the database and user browsers and all the important data access should be encrypted and stored in the back end of the firewall. PHDA also arranges professional management personnel to monitor the various processes of data collection, storage and access in real time to prevent users from malicious operations.

[1] Annex XVII: Network security situation monitoring record, including the monitoring of host situation and attack intrusion.

[2] Annex XII: Security monitoring and auditing system. This document describes the monitoring and auditing strategies for network security, host security, database security, application system security, etc.

[3] Annex XIII: Emergency plan. This document describes measures to be taken in the case of storage equipment failure, network equipment failure or server failure, natural disasters (flood, fire, electricity, etc.) and other emergencies (https://www.ncmi.cn/shareDocument/findContentManagementRulesDetail.do?id=2819).

[4] GB/T 22239-2019 Information security technology - Baseline for classified protection of cybersecurity, Chinese vision on http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=BAFB47E8874764186BDB7865E8344DAF

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These Requirements are not seen as final, and we value your input to improve the CoreTrustSeal certification procedure. Any comments on the quality of the Requirements, their relevance to your organization, or any other contribution, will be considered as part of future iterations.

Response:

Reviewer Entry

Reviewer 1

Comments:
Much better this time, made a good effort to address all previous comments and provide references and modified a few webpages as suggested. I found that Apples translation to English worked pretty well and https://www.onlinedoctranslator.com/app/translationprocess-pdf worked well for some of the Chinese documents they linked to and to parts of the PDF appendix that were not images.

Reviewer 2

Comments:
This is a very good application with extensive evidence documented and made publicly available.

Note that if the repository is increasinglyl open to non domestic users (users and/or deposits from outside China), an English interface and translation of the documentation will become essential.

PHDA API seems to only support machine harvesting using a national standard and is still developing an API compliant with international standards. Until such API is available the harvesting is limited which decreases interoperability of PHDA with other systems. The reviewer encourages the repository to prioritize the development of the API compliant with international standards