

Population Health Data Archive (PHDA)

Management Process

(Document Version: V2.0)

Release Date: May 2025

National Population Health Data Center, China

1 Purpose and Scope

This document describes the data management processes of the Population Health Data Archive (PHDA) that is managed by the National Population Health Data Center (NPHDC). It outlines the technical strategies used by the PHDA in the short, medium and long-term preservation of population health scientific data, and how these processes are managed by the staffers of the NPHDC.

2 Data Flow

The construction of Population Health Data Archive (PHDA) refers to the reference model of open archive information system (OAIS) made by the Space System Advisory Committee of NASA. OAIS is an ISO standard, which provides a functional framework for storing digital objects in managed repositories and guides the data management processes from data registration to storage and distribution. The main processes of PHDA include six functional entities: data ingest, archival storage, data management, administration, preservation planning and data access. In the PHDA system, the archived contents interact in the form of information packages, including Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP). The data flow of information in the long-term preservation process is shown in Figure 1.

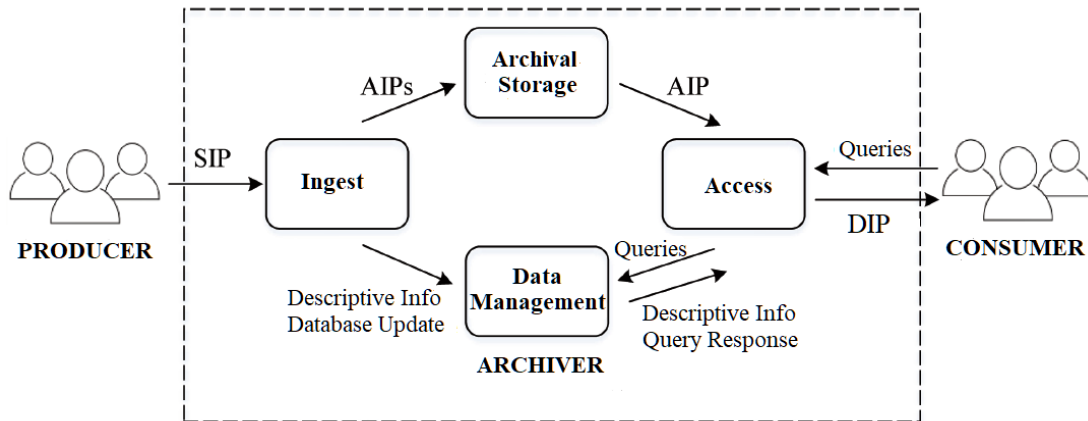


Figure 1 Data flow in the long-term preservation

In Figure 1, the data provider submits the content to be preserved to the PHDA system. SIP needs to include data and content information to ensure that the PHDA system can maintain the preserved content, and data users can access, understand and use the preserved content through the PHDA system.

The PHDA system receives SIP from the data provider, transforms SIP into a set of AIPs suitable for data archiving and data management through the data ingest function entity, classifies the received information objects, determines the set of each

object, and creates information to update the set description after AIP archiving.

The archival storage function entity receives the AIP generated by the ingestion process and adds it to the permanent repository. The data management functional entity uses the package description generated in the data ingestion phase, and extends the existing set description. In the processes of data archiving and management, media update, error correction and database maintenance are needed to prevent information loss caused by changes in technology, media, data format and user groups over time.

The data access function entity interacts with archival storage and data management according to the data access request of data users, and interacts with the AIP corresponding to DIP and its related information package description. Archival storage and data management create a copy of the request object in the temporary storage. Data access transforms the set of AIPs and associated package descriptions into a set of DIPs, and stores them in the physical dissemination media for delivery to data consumers in the data dissemination session.

3 Information Packages

The notion of OAIS information packages distinguishes the data format used in the exchanges between the PHDA and the external participants (producers, users). There are three kinds of information packages in PHDA.

3.1 Submission Information Package (SIP)

The Submission Information Package (SIP) is the data set which is provided by the producer. In the PHDA, the SIP consists of one or several data product files, metadata related to data processing, data dictionaries, data standards, proof of ethical review, software, toolkits, processing reports, data quality reports and other supporting documents. The metadata is initially provided by the producer or by NPHDC staffers in collaboration with the producer. This information is filled in by the producer, and checked and completed by authorized NPHDC staffers. The producer can edit the metadata by using the metadata generation tool and submit SIP to the data center for approval through the online registration system.

3.2 Archival Information Packet (AIP)

One or more SIPs are transformed into one or more Archival Information Package (AIP) for data release and preservation. In the PHDA, the AIP is the data set which is prepared by competent and authorized NPHDC staffers. The ingestion process and analysis builds new metadata, index and supporting files used by the metadata generation tool and archive production software in the PHDA. All contents will be organized into a standard directory structure. At the end of the ingestion process, the final AIP may contain the following items:

- (1) Complete set of data product files, includes auxiliary data files and calibration data files;
- (2) Complete metadata;
- (3) Software and tools related to data processing and application processing reports, data quality reports and other supporting documents.

The details of the AIP are shown in Figure 2, containing content information and preservation description information. Content information is the target information archived, which is composed of content data object and its related representation information, so that the content data object can be understood by the designated community. There are five types of preservation description information: reference information, provenance information, context information, fixity information and access information. The reference information provides a unique identifier to identify the content information. The provenance information describes the source of content information, provides the audit and tracking of content information, and provides the basis for the authenticity and reliability of content information. The context information records the reasons for creating the context information and its relationship with other content information objects in the context. The fixity information provides the inspection and verification of the data integrity for tracing the change record of the content information object. The access information provides the scope of permission to preserve, disseminate and use the content information.

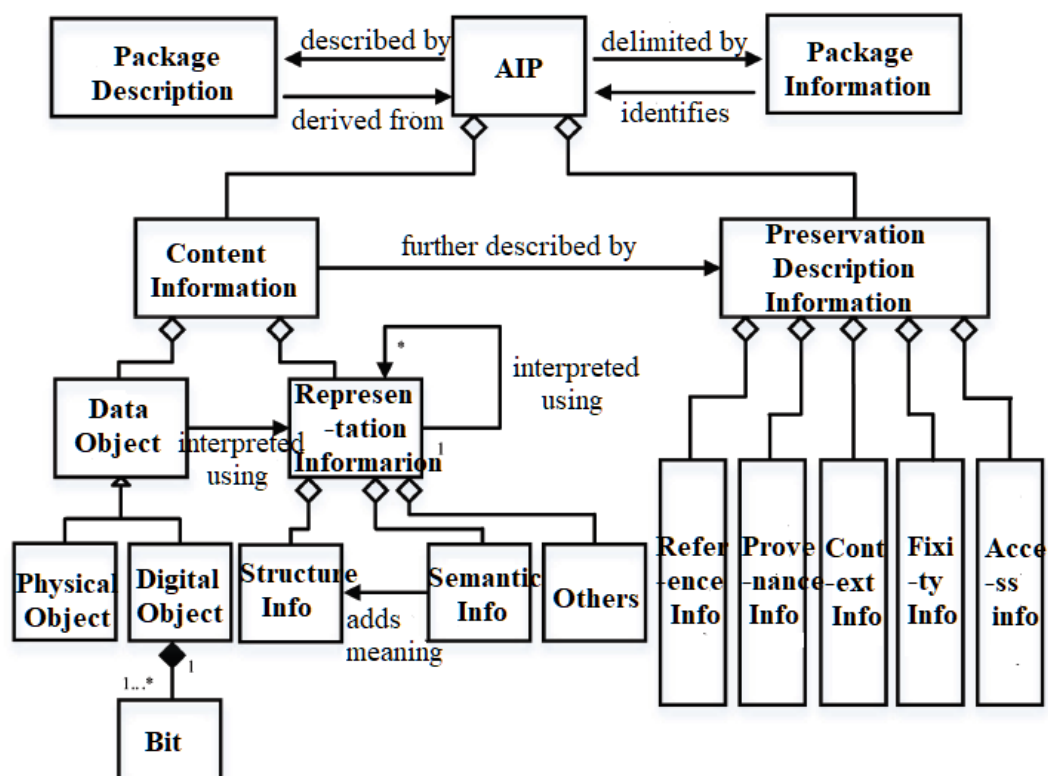


Figure 2 AIP Data Model

3.3 Dissemination Information Package (DIP)

The Dissemination Information Package (DIP) is the data set that NPHDC ultimately disseminates to the consumer. DIP may be all or part of AIP, even a set of several AIPs. The DIP may include the following items:

- (1) Data product files, includes auxiliary data files and calibration data files;
- (2) Metadata;
- (3) Software and tools related to data processing and application.

The PHDA data are made openly available to users through the data browsing function in the PHDA system (<https://www.ncmi.cn/phda/browse.html>). The interface allows for browsing, searching, filtering, and obtaining of the DIP.

4 Data Ingest

4.1 Ingest Process

The data ingest process mainly includes three parts: data receiving, data checking and preprocessing, and data objects ingest, as shown in Figure 3.

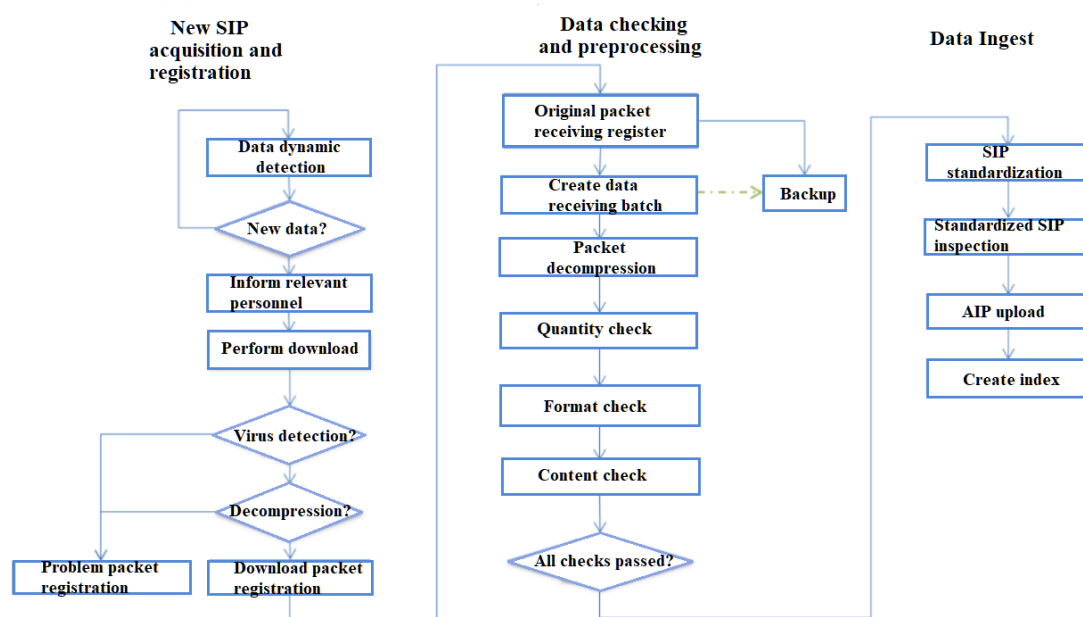


Figure 3 Data Ingest Pipeline

The data producer organizes data files and metadata into SIP, then submits data to NPHDC through the PHDA system. Data submission is divided into online and offline. Online submission includes new metadata, data upload (including but not limited to metadata, raw data, data dictionary or related instructions, instance data, data related papers, tools, etc.), certificate application, data and certificate audit, data release according to shared protocol, etc. Offline upload includes new metadata, filling in offline submission information, selecting data delivery mode (including but not limited to CD-ROM, mobile hard disk, U disk, etc.), data and certificate audit,

administrator uploading data, data release according to shared protocol, etc. The data producer is responsible for the quality of data content and metadata content, who also in cooperation with NPHDC, figures out the data quality problems happens in SIP submission and ingestion processes.

NPHDC performs virus detection and decompression on the received data, and then conducts quality audit. The quality audit mainly includes: whether the submitted scientific data is complete, whether it meets the form requirements, whether it meets the confidentiality requirements, whether it meets the specified quality requirements, whether it meets the requirements of non-sensitive information, whether it meets the requirements of non-ethical issues, whether the submitted scientific data has scientific value and use value, etc. There are two ways of quality audit: machine audit and manual audit, in which machine audit is mainly used for type, format, quantity, standardization and other data forms, while manual audit is mainly used for data content detection to ensure the quality of data obtained from various data sources. In addition, for formats that do not meet the requirements of long-term preservation, the necessary format conversion should be made on the basis of saving the original documents.

After data checking and preprocessing, the data center organizes the data reviewers to enhance the data, perfect and optimize the data set description information (including data classification, topic classification, keyword description accuracy, etc.), and enhance the description of data set metadata, such as filling in the theme words of Chinese Medical Subject Headings (CMeSH), expanding the description semantics of the data set, and adding data characteristic tags. Then the data center performs standardized operations on the data, generates metadata required by the preservation system (including description metadata, technical metadata, preservation metadata, etc.) according to the archive data (AIP) structure defined by the system, and classifies the received data objects to determine the collection of each object. Finally, the data is uploaded to the preservation system for AIP archiving, and the index is updated, backup is performed and ingest report is generated.

4.2 Membership Responsibilities

Responsibilities of members in SIP submission and ingestion are as follows:

(1) Data producers (person, research team or institutes) are responsible for the generation of data products and metadata, SIP organization in accordance with NPHDC SIP requirements and SIP submission into NPHDC. Data producers are also in charge of SIP data quality, ensuring the readability, specification, integrity, consistency, and availability of the data and metadata, figuring out the quality problems fed back by NPHDC in data ingest process.

(2) Data ingest specialist is responsible for SIP reception, SIP data quality

audition, examination (validate the automatically examination results and examine the SIPs manually further) and examination results feeding back. Data Ingest Specialist would organize peer experts to validate the precision and availability of data, organize the training and popularization of data model, metadata standards and technical system.

(3) Data archive specialist is responsible for the data set generation, designing the catalogs and attributes of the data set by considering different sub-discipline characteristics and storing the data set into database.

(4) Maintenance technicians are responsible for the operation, maintenance and upgrading of technical systems in NPHDC, such as data ingest web service, metadata editing tools, data archiving system, et al.

(5) Peer expert is responsible for data evaluation such as data generality, usability and accuracy.

5 Archival Storage

PHDA is composed of long-term storage system, data backup system and data disaster system.

5.1 Data Storage

Data storage, which determines the sensitive level of data (including extremely sensitive data, high sensitive data, medium sensitive data, low sensitive data and no sensitive data) according to the type, quantity, whether it involves user privacy and the degree of damage to the data, carries out hierarchical storage, and sets up security control measures (involving time, space, personnel and technology) to ensure the storage security of data in the PHDA. PHDA uses online-near line-offline multilevel storage. Online storage uses SAN storage and distributed storage, and near line and offline storage use tape libraries. SAN storage consists of EMC SAS array and switches, with a total capacity of about 1.5PB which is mainly used to support business cache and database cluster. Distributed storage is composed of DELL storage server and distributed file system. Distributed storage is divided into three categories: distribution cache, release cache and long-term archive storage, with capacities of 1PB, 1PB and 2.5PB respectively. The tape library uses the Quantum i6000 with a maximum capacity of 5000 tapes and 16 LTO9 drives to store inactive data in archival data. All the hardware and software in PHDA like server, disk array, tape and file system are equipped with corresponding management system, which can achieve real-time monitoring and maintenance.

5.2 Data Backup

NPHDC formulates data backup classification standard based on "Information

System Disaster Recovery Specification "(GB/T 20988-2007), adopts a 5-level backup strategy, and conducts differentiated management of various types of data. Data backup classification standard includes data backup media, backup data storage location, data transmission mode, data backup tool, data recovery point target (RPO), backup data source, data backup mode, backup data verification standard, backup data dump requirement, backup data archiving requirement, backup data deletion requirement, backup data recovery tool and other elements, as detailed in Table 1.

Table 1 Data backup classification standard

Data Backup Level	Level 1 local backup	Level 2 basic remote backup	Level 3 regular network backup	Level 4 real-time remote backup	Level 5 real-time synchronous backup
Data backup media	Disk, tape or CD	Tape or CD	Disk or tape	Disk	Disk
Backup data storage location	Local	Same city different places	Same city different places or remote different places	Remote different places	Same city different places
Data transmission mode	No transmission or LAN transmission	Transportation or MAN transmission	MAN transmission or WAN transmission	High-speed WAN transmission	Metro private storage network transmission
Data backup method	Manual or automated tools	Manual or automated tools	Manual or automated tools	Automation tools	Automation tools
Data recovery point target	0-2 weeks (inclusive)	1 day-1 week (inclusive)	2 hours -1 day (inclusive)	0-2 hours (inclusive)	0
Backup data sources	Running information system	Dumped data of Level 1 Local Backup or running information system	I Running information system	Running information system n	Running information system
Data backup mode	Real-time synchronous or real-time asynchronous or regular full backup or regular incremental	Full backup or incremental backup or selective backup	Full backup or incremental backup or selective backup	Real-time asynchronous	Real-time synchronization

	backups or regular selective backup				
Backup data validation standards	Sampling or indirect validation	Sampling or indirect validation	Sampling or indirect validation	Sampling validation	Sampling validation
Backup data dump requirements	Regular dump or no dump	Regular dump	Regular dump	No dump	No dump
Backup data archiving requirements	Regular archiving	Regular archiving	Regular archiving	No archiving	No archiving
Backup data deletion requirements	Regular recycling or medium demagnetizati on or physical destruction	Regular recycling or medium demagnetizati on or physical destruction	Regular recycling or medium demagnetizati on	Permanent reservations	Permanent reservations
Backup data recovery mode	Manual	Manual	Manual	Manual or automated tools	Manual or automated tools
Correspon den ce with GB/T20988 disaster recovery capability levels	No	Levels 1 and 2	Levels 3 and 4	Level 5	Level 6

The data backup methods include the following three kinds: (1) Full backup, the operation of fully backing up the host data is relatively time-consuming. Automatic backup is performed every Saturday (starting time 12:00). (2) Incremental backup, which only backs up the changed part of the data after the last backup. Compared with full backup, it takes less time. Incremental backup every Sunday to Friday (starting time 20:00) is triggered by the administrator or automatically executed by the system. (3) Selective backup, according to specific business requirements, the operation of backing up specific or critical data is more flexible than full and incremental backup, and it can be operated from time to time according to specific needs.

After PHDA generates data entities and data sets, it uses robotic arms and drives to automate backup and recovery, reducing manual operations and changing the

frequency of tapes, reducing human operating errors, and increasing the reliability of backups. Before backing up data, it automatically performs virus scanning and data verification on files to ensure that the backup data stored on the tape media is accurate. During disaster preparedness and reconstruction, if the system is damaged and the operating system and applications need to be reinstalled, the pre-installation operation can be skipped, and the damaged system and data can be restored to the original state in the shortest possible time. In addition, the data center will periodically check the MD5 value of the data stored locally and remotely to ensure data consistency.

5.3 Data Recovery

When data service specialist or data archive specialist finds occurrence of data loss, damage or not accessible, they will inform the maintenance technician to check the system status. After the technician confirms the abnormal data, they will activate recovery task and PHDA will execute operation of moving tape data or recovering the disaster data automatically.

5.4 Membership Responsibilities

The responsibilities of the members in the archival storage process are as follows:

(1) Data ingestion specialist is responsible for the initial review of the integrity of the SIP submitted by the producer.

(2) Data archive specialist is responsible for restructuring directory and producing AIP after supplementary information according to discipline or specification.

(3) Data service specialist is responsible for composing AIP to SIP which is demanded by consumer.

(4) Maintenance technician is responsible for monitoring and upgrading the software and hardware of PHDA and performing data storage, backup and recovery operations.

(5) Peer experts provide professional advice for PHDA's archive, storage, backup, etc.

6. Data Management

6.1 Management Functions

Data management functions include managing archive databases, performing database updates, providing query functions to manage data to produce result sets, and generating related reports from these retrieval result sets.

PHDA manages and retrieves AIP and metadata through product database and

archive database. After AIP preparation, the information of data product files and data labels is stored in the product database. The organization information and metadata information of data set are stored in the archive product database. All the information is managed and stored uniformly.

Data archive specialist copies, queries, analyses, transforms and checks data product files and data sets through product and archive management software. As the content or version of data and metadata is updated, the data archive specialist will update the data product file information, metadata, data description documents, etc. in the product database and archive database.

PHDA performs real-time or scheduled backup of database tables, indexes, views and other information. Real time backup is realized by the database backup tools and the scheduled backup is realized by the backup script (expdb, mysqldump) written by the maintenance technician. The backup frequency adopts real-time incremental backup, daily full backup and weekly full backup according to the increment and importance of data.

6.2 Membership Responsibilities

The responsibilities of the data management staff are as follows:

(1) Data archive specialists are responsible for metadata and AIP management, such as data update, long-term storage, disaster recovery, statistical analysis of archived data, regular inspection of data and metadata to ensure data security, usability and long-term accessibility.

(2) Maintenance technicians maintain information systems and application software, such as data product library, archive database, tape library, etc. to ensure the safe and stable operation of the system.

7 Data Access

7.1 Data Release

To ensure data security, PHDA provide data access functions and services based on the release database, a mirror of the archival database, rather than directly on the archival database. The data service specialists are responsible for the administration of the release database.

PHDA publishes data set metadata, sample data, data dictionaries, data entities 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 browsing, and provides version change and traceability of data, copyright and license of data, and reference and access of data.

7.2 Data Retrieval and Access

NPHDC provides public service-oriented medical digital resource distribution function for users within the scope of the agreement, and realizes the collaborative service of population health science data resources, including cross-system, cross-institution, cross-platform integration of resources, services, personnel, and technology. It provides users with one-stop service from data retrieval to acquisition, and supports browsing, retrieval, download, authority control, malicious download detection, user use statistics and other functions. The following data retrieval and access services are available at any time to PHDA consumers:

(1) Online data access service: PHDA portal (<https://www.ncmi.cn>) provides 7*24 hours' online data access service. Users can retrieve, browse or download data through the website, and check additional information such as metadata, unique identifiers, data usage licenses and so on online. The updated information of the released data will also be publicized on the website. PHDA have been registered on the re3data.org to promote findability and accessibility of NPHDC archive data. The access data are divided into open data, controlled data and closed data. Users can access the open data in the database, but must be approved to access the controlled data in the database. Data users can not infringe the rights of any individual or entity, and PHDA does not grant users ownership or intellectual property of any content, data, information that may be accessed in this warehouse. For clearly stated data, the data user has the responsibility to obtain the permission of the obligee before using the relevant data. For higher shared level scientific data, PHDA provides remote virtual desktop, and users can access and analyze remote data. If the data user violates the use specification of controlled or public data, PHDA will take appropriate measures.

(2) Cooperative access services: PHDA metadata can also be obtained through the data directory on the China Science and Technology Resource Sharing Network (<https://escience.org.cn/>). Users can find and access NPHDC data through China Science and Technology Resource Sharing Network.

7.3 Recovery Service

Release databases operate in dual-master mode to ensure high availability, and use database tool or scripts to implement daily backup. PHDA website adopts active-standby deployment by using KeepAlive services. When one host goes down, the service can be automatically switched to another host to ensure the normal external service of PHDA portal and retrieval system.

7.4 Membership Responsibility

The responsibilities of data access staff include:

(1) The data service specialist is responsible for tracking user requirements, answering user questions, processing user orders and providing data products to users.

(2) Maintenance technicians maintain information systems and application software, such as publishing database, large file transfer tools and PHDA website, etc. to ensure the safe and stable operation of the system.

8 Administration

8.1 Management Functions

Administration provides services for the overall operation of the PHDA, including system management configuration, access control, user behavior monitoring, ingest policy formulation, system operation monitoring, logging and other functions to achieve data preservation life cycle management.

PHDA system consists of hardware and software components. The hardware components are the infrastructures include server, storage, network, firewall, etc. The software components functions include data acquisition and ingest, data processing pipeline, data quality control, data management, data archive storage and data access services. NPHDC has formulated a detailed system management configuration scheme, and deployed a monitoring platform for unified monitoring and management of the PHDA system hardware equipment. The data center uses the log audit system to comprehensively collect the logs (including operation, alarm, message, status, etc.) generated by the security equipment, network equipment, databases, servers, hosts and so on, and perform store, monitoring, auditing, analysis, alarm, response and report.

PHDA system implements strict access control, sets user permissions, and adopts dual-factor identity authentication, and centrally manages the operation and maintenance personnel through VPN + fortress machine. The system monitors the abnormal conditions in the access behavior according to the access control policy, such as unauthorized users operating data or attempting to change the system configuration, users downloading maliciously, etc., records the abnormal log and sends out an alarm. The system monitors the response result and response time of each user request. If a wide range of acquisition failures or time delays occur, the administrator should be notified immediately.

PHDA system records user access, login, registration, transfer data, download data and other operational behavior information, and form a detailed log for statistical analysis. The information related to the ingest behavior includes the operator, time, ingest entry, success/failure entry, backup location, etc., and the ingest report is sent. The information of user access behavior includes user name, time, behavior, acquired information, whether it is successful, IP address, etc.

8.2 Management Activities

The operation, maintenance and evolution of PHDA system include the following management activities:

(1) The PHDA technical and management team (data ingestion specialist, data archive specialist, maintenance technicians, researcher, etc.) holds weekly meetings to discuss data set management issues.

(2) NPHDC convenes all staff and hold regular follow-up meetings within the data center.

(3) NPHDC staff keep in touch with data producers.

(4) The Council is convened annually to provide high-level guidelines and recommendations for PHDA development.

(5) NPHDC provides contact telephone number (010-65250541), email address (rkjkpt@126.com), and technical support telephone number (010-52328766) to help users use PHDA services, and provides feedback on content and access failures or improvement suggestions. NPHDC is responsible for reporting and answering callers in a short time. Users can contact data ingestion specialists in the data submission process.

8.3 Membership responsibilities

NPHDC sets up different roles for the PHDA operation and maintenance, including data ingestion specialist, data archive specialist, data service specialist, peer experts, maintenance technician, etc. Each role needs to carry out operation and maintenance activities according to the corresponding workflow and operation specification, the following activities needs to be done by the operating personnel.

(1) The events found during operation shall be notified to the relevant, and dealt according to the operating procedure and failure plan.

(2) Monitoring and routine maintenance should be done according to operation plan.

(3) Record the events of the operation and maintenance activities.

(4) Collect and archive operational and maintenance documents.

(5) Carry out analysis and evaluation regularly, and put forward suggestions for optimization and improvement.

9 Preservation Planning

NPHDC plans to make preservation and long-term preservation plans for data archives, take a series of technical measures for data management and information system management, and keep these measures updated to prevent the risk of obsolescence and ensure the long-term availability of data.

9.1 Data Sustainability

NPHDC implements data management based on the data model. In terms of data files and metadata, NPHDC refers to field or industry standards, tracks changes in standards. If necessary, the version of data files and metadata will be upgraded to ensure the sustainable use of data.

(1) Metadata management

NPHDC formulates the metadata management strategy, which not only retains the digital object itself, but also collects, enriches and processes various forms of metadata, including descriptive, technical, managerial, structural, permission management, process relevance and so on. Metadata files are stored in XML format. When metadata standards change, it is very easy to upgrade and transform. This method ensures the long-term interpretability, retrievability and interoperability of relevant data.

(2) Format management

The sustainability of data resource storage format is directly related to the authenticity, integrity and reliability of resources, and is the most critical issue in data resource storage and long-term preservation. The data center guarantees format sustainability through standardized format identification, validation and risk assessment.

A. Format recognition refers to the identification and description of the technical attributes of digital archives resources, including the generation environment, which generally exists in the form of technical metadata. When NPHDC receives data resources, the format of data is recognized, and the technical metadata is extracted and archived.

B. Format validation is to verify the validity of the format by verifying whether the existing features of the format conform to the expected features of the format. When NPHDC receives data resources, once the resource format is identified, it is necessary to verify its validity.

C. Risk assessment is to analyze the risk of format obsolescence, including the risk of format shutdown or format unreadable. NPHDC will convert all data resources in risk formats into standardized formats, and migrate information to new sustainable formats. NPHDC supports multiple data formats, in order to ensure the long-term accessibility, readability and usability of the data, the recommended formats for long-term data preservation are listed, as shown in Table 2.

Table 2 Recommended preservation formats for research data

Type of data	File formats suitable for digital preservation	Standard, widespread file formats	Examples of sources and applications
--------------	--	-----------------------------------	--------------------------------------

Audio	WAVE (*.wav), AIFF (*.aiff, *.aif), MXF (*.mxf), Matroska (*.mka)	WAVE(*.wav), MP3(*.mp3), AAC(*.aac, *.m4a, mp4), AIFF(*.aiff, *.aif), OGG(*.ogg), BWF(*.bwf), MXF(*.mxf), Matroska(*.mka), FLAC(*.flac), OPUS(*.opus)	Interviews, surveys
Biomaterial data	CSV (*.csv), TXT (*.txt), XML(*.xml)	CSV (*.csv), TXT (*.txt), FASTA (*.fasta), FASTQ (*.fq, *.fastq), PDB(*.pdb, *.ent, *.brk), XLS (*.xls), XML (*.xml)	DNA sequencers, mass spectrometers, microarrays, spectrophotometers
Classifications, thesauri, codes	PDF/A (*.pdf), XML (*.xml)	DOC (*.doc, *.docx), PDF (*.pdf), XML (*.xml)	Medical lexicon, ontology
Databases	SQL (*.sql)	SQL (*.sql), SIARD (*.siard), CSV (*.csv), MS Access (*.mdb, *.accdb), dBase (*.dbf), HDF5 (*.hdf5, *.he5, *.h5)	Biomedical database
Geospatial data	GML (*.gml), MIF/MID (*.mif/ *.mid)	GML (*.gml), MIF/MID (*.mif/ *.mid), ESRI Shapefiles (*.shp), MapInfo (*.tab), KML (*.kml)	Vector and raster data
Image data	TIFF (*.tif, *.tiff), PNG (*.png), JPEG2000 (*.jp2), SVG (*.svg)	TIFF (*.tif, *.tiff), JPG (*.jpg, *.jpeg), PNG (*.png), GIF (*.gif), JPEG 2000 (*.jp2), STL (*.stl), OPG (orthopantomogram), DICOM (*.dcm), DVT (digital volume tomography), PDF (*.pdf), SVG (*.svg), Illustrator (*.ai), EPS (*.eps)	Cameras, microscopes, MRT and CT scans, ultrasonic, X-ray and sonography instruments
Image data 3D	OBJ (*.obj, *.mod in ASCII format), X3D (*.x3d), VRML (*.vrml, *.wrl)	STL (*.stl), OBJ (*.obj, *.mod), PLY (*.ply), DXF (*.dxf), X3D (*.x3d), COLLADA (*.dae), FBX (*.fbx), VRML (*.vrml, *.wrl)	3D technologies such as stereolithography
Markup language	XML (*.xml)	XML (*.xml), HTML (*.html), SGML (*.sgml)	Metadata

Sensor data	CSV (*.csv), TXT (*.txt), PDF (*.pdf)	CSV (*.csv), TXT (*.txt), PDF (*.pdf), XLS (*.xls, *.xlsx), XML (*.xml)	Thermal sensors, pressure sensors, polysomnography, ECG, EEG
Spreadsheets	CSV (*.csv)	ODS (*.ods, *.odt, *.odg, *.odc, *.odf), CSV (*.csv), XLS (*.xls, *.xlsx), OOXML (*.docx, *.docm), PDF/A (*.pdf)	Data from research, clinical care
Statistical data	R (*.r), CSV (*.csv)	CSV (*.csv), SAS (*.7dat, *.sd2, *.tpt), SPSS (*.sav), SPSS Portable (*.por), STATA (*.dta), DDI (*.xml), data (*.csv, *.txt), R (*.r)	Data from research, clinical care
Text files	PDF/A (*.pdf), XML (*.xml), Unicode TXT (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py etc. in ASCII format)	DOC (*.doc, *.docx), PDF (*.pdf), TXT (*.txt), RTF (*.rtf), ODT (*.odt), Powerpoint (*.ppt)	Documentations, reports, findings, administrative data
Video	Matroska (*.mka), MXF (*.mxf), AVI (*.avi)	MPEG-4 (*.mp4, *.m4a, *.m4v), MPEG-2 (*.mpg, *.mpeg, *.m2v, *.mpg2), AVI (*.avi), MXF (*.mxf), Matroska (*.mka, *.mkv), QuickTime (*.mov, *.qt), Windows Media (*.wmv)	Cameras, CT scans, ultrasonic instruments

9.2 Data Integrity

NPHDC ensures data integrity by checking, redundant storage and timely replacing obsolete storage media. It mainly includes the following measures:

(1) Before receiving data resources, first scan them to ensure that no malware is entrained and handed over, and repair and check the data resources to verify whether the receiving state is consistent with the original state or whether it has been changed.

(2) Catalog all archived data resources, and record the receiving process of all resources and other life cycle processes after storage, including format migration, copying, and auditing and other long-term preservation processes.

(3) Replicate and back up data resources, especially those storage media that are at risk of obsolescence or destruction. If necessary, blockers can be used to prevent damage to digital content caused by physical media damage.

(4) Carry out annual spot checks on preserved data resources, including carrier spot checks and repair checks, and carry out carrier replacement or repair of defective resources.

(5) The digital contents of all storage carriers must be stored on new reliable carriers until the tenth year.

(6) Conduct a quarterly log review of the PHDA system to verify the integrity of the data resource content, structure, background information and its management process traceability.

9.3 Service Continuity

The development of IT information technology and changes in data user needs make PHDA need to track and maintain sustainability. NPHDC will collect system transformation requirements and upgrade the system through online user feedback mechanisms, consultation of subject experts and IT technical experts. System upgrades usually take place every 3-5 years. The upgrade of PHDA will be conducted by professional software companies in accordance with the requirements of software engineering management. Before the new version of PHDA is launched, system testing will be conducted to ensure the safety and reliability of the system. When the new version of the system is deployed, it will be implemented according to the migration plan to ensure the integrity and consistency of data and related information.

(1) Fund guarantee and support from affiliate institutions

The National Population Health Data Center (NPHDC) is one of the 20 national science data centers recognized by the Ministry of Science and Technology and the Ministry of Finance. It is a science and technology resource sharing service platform under the National Science and Technology Infrastructure Platform, and its supporting unit is the Chinese Academy of Medical Sciences. NPHDC construction funds are coordinated by the Ministry of Science and Technology and allocated by the Ministry of Finance. The fixed operating funds of the Ministry of Science and Technology can be obtained every year. Most of the NPHDC staff are staff members of the Chinese Academy of Medical Sciences, which can effectively ensure the stability of the staff. In addition, the Chinese Academy of Medical Sciences also provides Class A computer rooms and fixed workplaces for NPHDC. The above supporting conditions can ensure that NPHDC continues to provide data access services.

(2) Tracking and assessing trends in data management standards and data specifications

NPHDC established a dynamic monitoring system to track the development trends and new requirements of frontier data policies, standards and management requirements at home and abroad. These tracking results provide important references for PHDA related standards (such as AIP, SIP requirements) and data policy. The NPHDC tracking list is as follows:

A. Classification and coding standard of population health science data sharing dataset

- B. Common data formats
- C. GB/T 32843-2016 Science and technology resource identification
- D. GB/T 36344-2018 Information technology—Evaluation indicators for data quality
- E. GB/T 30523 Core metadata of science and technology resources
- F. GB/T 31073 Core metadata of science and technology resources service

.....

(3) Tracking and evaluating the development trend of the technology system

NPHDC information system construction refers to or follows the following national standards and international standards. NPHDC conducts long-term follow-up on the development of the following standards and evaluates technical system upgrade needs based on the follow-up results.

- A. ISO 14721 Reference Model for an Open Archive Information System
- B. ISO 16363 Audit and certification of trustworthy digital repositories
- C. ISO 20652-2006 Producer-archive interface — Methodology abstract standard
- D. GB/T 25058-2010 Information security technology—Implementation guide for classified protection of information system
- E. GB/T 22239-2019 Information security technology—Baseline for classified protection of cybersecurity
- F. GB 4943.1-2011 Information technology equipment Safety Part1: General requirements
- G. GB 4943.23-2012 Information technology equipment safety Part23: Large data storage equipment
- H. GJB9001C Quality management systems requirements

.....

(4) Tracking user service requirements and product technologies

In order to improve services, NPHDC defines the following ways to track the service requirements and product technologies of users (producer and consumer).

A. Tracking requirements during Data Archive Plan preparation. Through the preparation process of the Data Archive Plan, the NPHDC 's archive specialist negotiates with the producer on the format, version, dedicated software tools for archiving data, archive arrangement, and archive interface, etc. The archive specialist also confirms these contents during each archiving process. This process is one of the effective ways for NPHDC to track data producer's service and product technology needs.

人口健康科学数据仓储（PHDA） 管理流程

(文档版本：V1.0)

国家人口健康科学数据中心

发布日期：2020 年 10 月 10 日

1 目的与范围

本文件描述了由国家人口健康科学数据中心(NPHDC)管理的人口健康科学数据仓储(PHDA)的数据管理流程。概述了 PHDA 在人口健康科学数据短、中、长期保存各个过程采用的技术策略，以及 NPHDC 工作人员如何管理这些过程。

2 数据流程

人口健康科学数据仓储（PHDA）建设参考美国航天局空间系统咨询委员会制作的开放档案信息系统（OAIS）参考模型。OAIS 是一个 ISO 标准，提供了用于在托管存储库中保存数字对象的功能框架,指导了数据注册获取到存储分发的数据管理过程。PHDA 的主要流程包括数据摄入、数据存档、数据管理、系统管理、保存规划和数据访问等六个功能实体。在 PHDA 系统中，保存内容以信息包的形式进行交互，包括提交信息包(Submission Information Package, SIP)、存档信息包 (Archival Information Package , AIP) 和分发信息包 (Dissemination Information Package, DIP)，保存信息在长期保存过程中的数据流向见图 1。

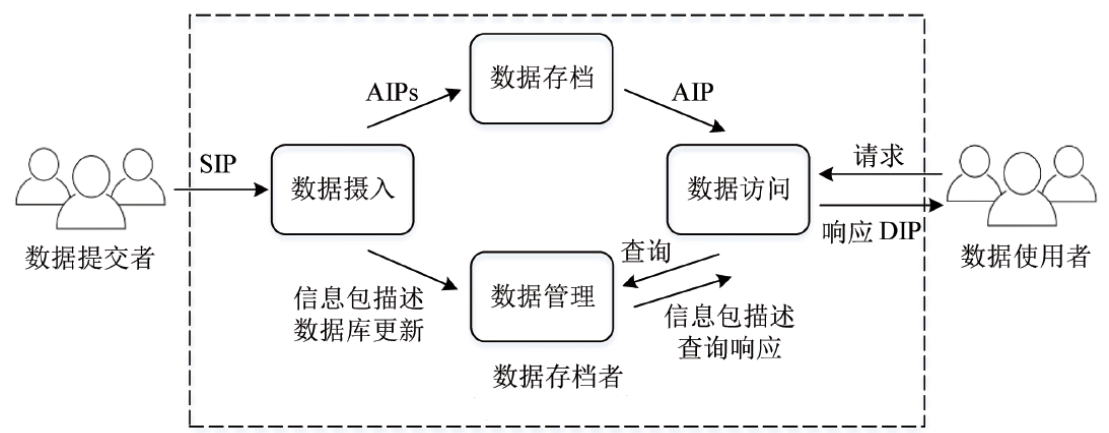


图 1 长期保存过程中的数据流向

图 1 中，数据提交者向 PHDA 系统提交将要保存的内容，SIP 需要包含数据及内容信息，以确保 PHDA 系统可以维护保存内容，数据使用者可以通过 PHDA 系统访问、理解和使用保存内容。

PHDA 系统接收来自数据提交者的 SIP，通过数据摄入功能实体将 SIP 转换为适合于数据存档和数据管理所需的一组 AIP，并对接收的信息对象进行分类，确定每个对象的所属集合，完成 AIP 存档后创建消息以更新集合描述。

数据存档功能实体接收摄入流程生成的 AIP，并将其添加到永久保存库。数据管理功能实体采用数据摄入阶段生成的包描述，并扩展现有的集合描述。数据在存档和管理过程中需要进行媒介更新、纠错以及数据库维护等操作，以防技术、媒介、数据格式和用户群体等随着时间推移所产生的变化而导致信息丢失。

数据访问功能实体根据数据使用者的数据访问请求，与数据存档和数据管理相互作用，交互 DIP 所对应的 AIP 及其相关信息包描述。数据存档和数据管理

在临时存储中创建请求对象的副本，数据访问将该组 AIP 和关联的包描述转换为一组 DIP，并将其存储于物理分发媒介，以在数据分发会话中交付给数据使用者。

3 信息包

OAIS 信息包的概念区分了 PHDA 与外部参与者(数据生产者、数据使用者)之间交流时使用的数据格式。在 OAIS 中有三种信息包。

3.1 提交信息包(SIP)

提交信息包(SIP)是生产者提供的数据集。在 PHDA 中，SIP 由一个或多个数据产品文件、与数据处理相关的元数据、数据字典、数据标准、伦理审查证明、软件、工具包、处理报告、数据质量报告和其他支持文档组成。元数据由生产者提供，或由 NPHDC 工作人员与生产者协作提供。这些信息由生产者填写，由授权的 NPHDC 工作人员检查和完善。生产者可以使用元数据生成工具编辑元数据，并通过在线注册系统将 SIP 提交给数据中心审批。

3.2 存档信息包(AIP)

将一个或多个 SIP 转换为一个或多个存档信息包(AIP)，用于数据发布和保存。在 PHDA 中，AIP 是由授权的 NPHDC 工作人员准备的数据集。摄入过程构建了新的元数据、索引和支持文件，这些文件供 PHDA 中的元数据生成工具和档案制作软件使用。所有内容被组织成一个标准的目录结构。在摄入过程结束时，最终的 AIP 可能包含以下内容：

- (1) 全套数据产品文件，包括辅助数据文件和校验数据文件等；
- (2) 完整的元数据；
- (3) 与数据处理和应用处理报告、数据质量报告及其他支持文件相关的软件和工具。

AIP 详细结构见图 2，包含内容信息和保存描述信息。内容信息是保存的目标信息，由内容数据对象及其相关的呈现信息组成，以使内容数据对象可被指定团体理解；保存描述信息有指引信息、起源信息、环境信息、不变性信息以及访问权限信息 5 种类型。其中，指引信息提供唯一标识符，以标识内容信息；起源信息描述内容信息的来源，提供对内容信息的审核跟踪，为内容信息的真实性和可靠性提供依据；环境信息记录创建内容信息的原因及其与环境中的其他内容信息对象之间的关系；不变性信息提供对数据完整性的检查与验证，用于追溯内容信息对象的更改记录；访问权限信息提供对保存、分发和使用内容信息的权限许可范围。

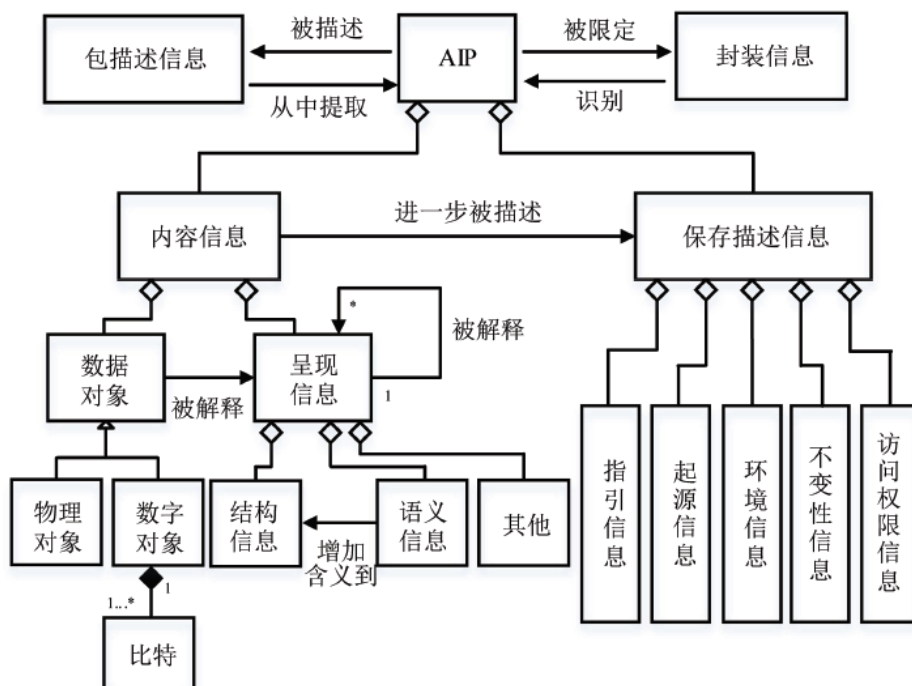


图 2 存档信息包数据模型

3.3 分发信息包(DIP)

分发信息包(DIP)是 PHDA 最终分发给数据使用者的数据集。DIP 可以是 AIP 的全部或部分，也可以是一组多个 AIP。DIP 可能包括以下内容：

- (1) 数据产品文件，包括辅助数据文件和校验数据文件；
- (2) 元数据；
- (3) 与数据处理和应用处理报告、数据质量报告及其他支持文件相关的软件和工具。

PHDA 数据通过 PHDA 系统数据浏览功能向用户开放 (<https://www.ncmi.cn/phda/browse.html>)。该界面允许用户浏览、检索、筛选以及获取数据。

4 数据摄入

4.1 摄入流程

数据资源的摄入过程主要包括数据接收、数据检查与预处理、数据对象的摄入三个部分，见图 3。

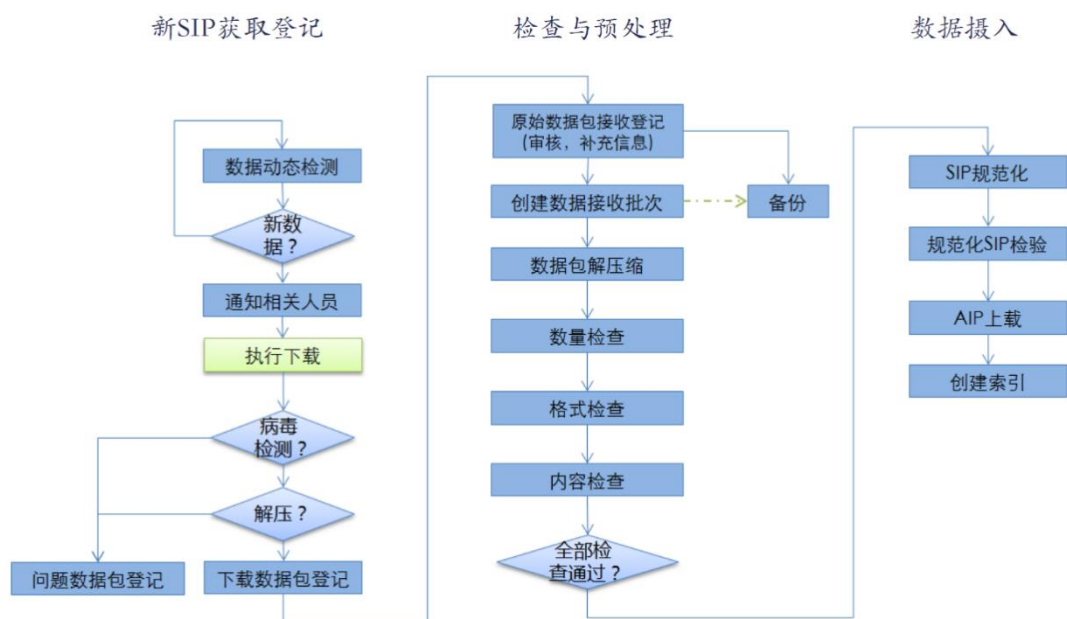


图3 数据摄入流程

数据提交者将数据文件和元数据组织成 SIP，通过 PHDA 系统进行数据提交。数据提交分为线上或线下两种方式。线上提交包括新建元数据、数据上传（包括但不限于元数据、原始数据、数据字典或相关说明、实例数据、数据相关论文、工具等）、证书申请、数据与证书审核、数据按共享协议发布等环节；线下上传包括新建元数据、填写线下提交信息、选择数据递送方式（包括但不限于光盘、移动硬盘、U 盘等形式的存储）、数据与证书审核、管理员上传数据、数据按共享协议发布等环节。数据生产者负责数据内容和元数据内容的质量，并与 NPHDC 合作，指出数据 SIP 提交和摄入过程中发生的数据质量问题。

NPHDC 对接收的数据进行病毒检测和解压缩，然后进行质量审核。质量审核主要包括：审核提交的科学数据是否齐全、是否符合形式要求、是否符合保密要求、是否符合规定的质量要求、是否符合无敏感信息要求、是否符合无伦理问题要求、提交的科学数据是否具有科学价值和使用价值等。质量审核途径包括机器审核及人工审核两种，其中机器审核主要用于类型、格式、数量、规范性等数据形式检测，而人工审核则主要用于数据内容检测，保证从各数据源获取的数据质量。此外，对于不满足长期保存要求的格式，应在保存原有文件的基础上做必要的格式转换。

数据检查与预处理完成后，数据中心组织数据审编员进行数据增强，对数据集描述信息（包括数据分类、主题分类、关键词描述准确性等）进行完善和优化，对数据集元数据进行增强描述，如补充中文医学主题词表 (Chinese Medical Subject Headings, CMeSH) 的主题词标注，扩展数据集描述语义，增加数据特色标签。而后数据中心对数据进行规范化操作，根据系统定义的存档数据（AIP）结构，生成保存系统需要的元数据（包括描述元数据、技术元数据、保存元数据等），并对接收的数据对象进行分类，确定每个对象的所属集合。最后将数据上载到保

存系统进行 AIP 存档，并更新索引、执行备份和生成摄入报告。

4.2 成员责任

成员在提交和摄入 SIPs 时的责任如下：

(1) 数据生产者(个人、研究团队或研究机构)负责数据产品和元数据的生成，按照 NPHDC 的要求组织 SIPs，并提交到 NPHDC。数据生产者负责 SIPs 数据质量，确保数据和元数据的可读性、规范性、完整性、一致性和可用性，解决 NPHDC 反馈的数据摄入过程中的质量问题。

(2) 数据摄入管理员负责 SIPs 接收、SIPs 数据质量审核、检测(自动验证审核结果并进一步手动审核 SIPs)、审核结果反馈。数据摄入管理员组织同行专家验证数据的准确性和可用性，组织数据模型、元数据标准和技术系统的培训和普及。

(3) 数据存档管理员负责数据集的生成，根据不同的子学科特点设计数据集的目录和属性，并将数据集存储到数据库中。

(4) 运维工程师负责系统的运行、维护和升级 NPHDC 的技术系统，如数据摄入 WEB 服务、元数据编辑工具、数据归档系统等。

(5) 同行专家负责数据评估，如评估数据的通用性、可用性和准确性。

5 数据存档

PHDA 由数据存储系统、数据备份系统和数据灾备系统组成。

5.1 数据存储

数据存储，根据人口健康科学数据的类型、数量、是否涉及用户隐私等敏感信息以及数据遭到破坏后的危害程度等确定数据的敏感级别（包括极敏感数据、高敏感数据、中敏感数据、低敏感数据和无敏感数据，共 5 级），进行分级存储，并设置对应环节的安全管控措施(涉及时间、空间、人员和技术)，保障数据在 PHDA 中的存储安全。PHDA 使用在线-近线-离线多级存储。在线存储使用 SAN 存储和分布式存储，近线存储和离线存储使用磁带库。SAN 存储由 EMC SAS 磁盘阵列和交换机组成，总容量约 500TB，主要用于支持业务缓存和数据库集群。分布式存储由戴尔存储服务器和分布式文件系统组成。分布式存储分为分发缓存、发布缓存和长期档案存储三类，容量分别为 800TB、800TB 和 2PB。磁带库使用最大容量为 3000 个磁带和 16 个 LTO6 驱动器的 Quatumn i6000 来存储归档数据中的非活动数据。PHDA 中所有的硬件和软件，如服务器、磁盘阵列、磁带、文件系统都配备了相应的管理系统，可以实现实时监控和维护。

5.2 数据备份

NPHDC 基于《信息系统灾难恢复规范》(GB/T 20988-2007) 制定数据备份分级标准，采用 5 级备份策略，对各类数据进行差别化管理。数据备份分级标准包含数据备份介质、备份数据存放地点、数据传输方式、数据备份工具、数据恢

复点目标（RPO）、备份数据来源、数据备份方式、备份数据验证标准、备份数据转储要求、备份数据归档要求、备份数据删除要求、备份数据恢复工具等要素，详见表 1。

表 1 数据备份分级标准

数据备份等级	第 1 级本地备份	第 2 级基本异地备份	第 3 级定期网络备份	第 4 级实时异地备份	第 5 级实时同步备份
数据备份介质	磁盘、磁带或光盘	磁带或光盘	磁盘或磁带	磁盘	磁盘
备份数据存放地点	本地	同城异地	同城异地或远程异地	远程异地	同城异地
数据传输方式	无传输或局域网传输	交通工具或城域网传输	城域网传输或广域网传输	高速广域网传输	城域专用存储网络传输
数据备份方式	人工或自动化工具	人工或自动化工具	人工或自动化工具	自动化工具	自动化工具
数据恢复点目标	0-2 周（含）	1 天-1 周（含）	2 小时-1 天（含）	0-2 小时（含）	0
备份数据来源	正在运行的信息系统	“第 1 级本地备份”的转储数据或正在运行的信息系统	正在运行的信息系统	正在运行的信息系统	正在运行的信息系统
数据备份方式	实时同步或实时异步或定期全量备份或定期增量备份或定期选择性备份	全量备份或增量备份或选择性备份	全量备份或增量备份或选择性备份	实时异步	实时同步
备份数据验证标准	抽样或间接验证	抽样或间接验证	抽样或间接验证	抽样验证	抽样验证
备份数据转储要求	定期转储或无转储	定期转储	定期转储	无转储	无转储
备份数据归档要求	定期归档	定期归档	定期归档	无归档	无归档

备份数据删除要求	定期循环使用 或介质消磁 或物理销毁	定期循环使用 或介质消磁 或物理销毁	定期循环使用 或介质消磁	永久保留	永久保留
备份数据恢复方式	人工	人工	人工	人工或自动化工具	人工或自动化工具
与 GB/T20988 灾难恢复能力等级的对应关系	无	第 1 级和第 2 级	第 3 级和第 4 级	第 5 级	第 6 级

数据备份方法包括如下三种：（1）全量备份，对主机数据进行全部备份的操作，耗时相对较多。每周六（开始时间 12: 00）进行自动备份。（2）增量备份，只备份上次备份后数据的改变量部分，相对全量备份而言，耗时较少。每周日至周五增量备份（开始时间 20: 00），由管理员触发实现或由系统自动执行。（3）选择性备份，根据具体业务需求，针对特定或关键数据进行备份的操作，相比全量和增量备份，灵活性更高。根据具体需求不定时进行操作。

PHDA 在生成数据实体和数据集后，使用机械臂和驱动器进行自动化备份与恢复，减少人工操作和更换磁带频率，降低人为操作失误，增加备份的可靠性。备份数据前自动对文件进行病毒扫描和数据校验，确保存储在磁带介质上的备份数据是准确的。灾难防备与重建时，可以在系统毁损且需重新安装操作系统、应用程序的情况下，跳过前期安装操作，直接将损毁的系统和数据在最短的时间内恢复原状。此外，数据中心将定期检查存储在本地和远程的数据的 MD5 值，以确保数据的一致性。

5.3 数据恢复

当数据服务管理员或数据存档管理员发现数据丢失、损坏或无法访问时，将通知运维工程师检查系统状态。运维工程师确认数据异常后，启动恢复任务，PHDA 自动执行移动磁带数据或恢复灾难数据的操作。

5.4 成员责任

数据存档过程中各成员的责任如下：

- （1）数据摄入管理员负责对数据生产者提交的 SIP 的完整性进行初步审核。
- （2）数据存档管理员负责重组目录，根据学科或规范补充信息后生成 AIP。
- （3）数据服务管理员负责按数据使用者需求将 AIP 与 SIP 进行组合。
- （4）运维工程师负责监控和升级 PHDA 的软硬件，执行数据存储、备份、恢复操作。

(5) 同行专家为 PHDA 的档案、存储、备份等提供专业建议。

6 数据管理

6.1 管理功能

数据管理功能包括管理存档数据库、执行数据库更新、对管理数据提供查询功能以产生结果集，并从这些检索结果集中生成相关报告。

PHDA 通过产品数据库和存档数据库来管理和检索 AIP 和元数据。AIP 准备后，数据产品文件信息和数据标签信息存储在产品数据库中。数据集的组织信息和元数据信息存储在归档产品数据库中。所有的信息都被统一管理和存储。

数据存档管理员通过产品和存档管理软件来复制、查询、分析、转换和检查数据产品文件和数据集。随着数据、元数据内容或版本的更新，数据存档管理员将更新产品数据库和存档数据库中的数据产品文件信息、元数据、数据描述文档等。

PHDA 对数据库表、索引、视图和其他信息进行实时或按计划备份。实时备份由数据库备份工具实现，按计划备份由运维工程师编写的备份脚本(expdb、mysqldump)实现。备份频率根据数据的增量和重要性，采用实时增量备份、每日全备份和每周全备份等方案。

6.2 成员责任

数据管理工作人员职责如下：

(1) 数据存档管理员负责元数据和 AIP 管理，如数据更新、长期存储、灾难恢复、存档数据统计分析、定期检查数据和元数据，确保数据的安全性、可用性和长期可访问性。

(2) 运维工程师负责维护信息系统和应用软件，如数据产品库、存档数据库、磁带库等，确保系统安全稳定运行。

7 数据访问

7.1 数据发布

为确保数据安全，PHDA 系统提供基于发布数据库的访问功能，发布数据库是存档数据库的镜像。数据服务管理员负责发布数据库的管理。

PHDA 根据用户授权和许可发布授权范围内的数据集元数据、样例数据、数据字典和数据实体及相关文档及工具。系统提供完整的、多维度的数据描述、展示和浏览，提供数据的版本变化和溯源功能、数据的版权和许可、数据的引用和访问方式。

7.2 数据检索和访问

NPHDC 对协议规定范围内的用户提供面向公共服务的医学数字资源分发功

能,实现人口健康科学数据资源协同服务,包括跨系统、跨机构、跨平台的资源、服务、人员及技术的集成整合交互共享,为用户提供从数据检索到获取的一站式服务,支持浏览、检索、下载、权限控制、恶意下载检测、用户使用统计等功能。

PHDA 用户可随时使用以下数据检索和访问服务:

(1) 在线数据访问服务: PHDA 门户网站(<https://www.ncmi.cn>)提供 7*24 小时的在线数据访问服务。用户可以通过网站检索、浏览或下载数据,并在线查看元数据、唯一标识符、数据使用许可等附加信息。发布数据的升级信息也将在网站上公示。PHDA 已经在 re3data.org 上注册,以促进 NPHDC 存档数据的可查找性和可访问性。

对于访问数据分为开放数据、受控数据和关闭数据。用户可以访问数据库中的开放数据,但是必须获得批准才能访问数据库中的受控数据。数据使用者不能侵犯任何个人或实体的权利,PHDA 并不授予用户在本仓储可能访问到任何内容、数据、资料的所有权或知识产权。对于明确权利声明的数据,数据使用者有责任取得权利人许可,方可使用相关数据。针对较高共享级别的科学数据,PHDA 提供远程虚拟桌面,用户可进行远程数据访问和数据分析。如果数据使用者违反受控或公开数据的使用规范,PHDA 将采取适当的措施。

(2) 合作访问服务: PHDA 元数据也可以通过中国科技资源共享网(<https://escience.org.cn/>)上的数据目录获得。用户可以通过中国科技资源共享网查找和访问 NPHDC 数据。

7.3 服务恢复

发布数据库以双主机模式部署,并使用数据库工具或脚本进行日常备份,保证高可用性和高可靠性。PHDA 网站系统通过使用 KeepAlive 服务进行主-备部署,当一台主机发生故障时,该服务可以自动切换到另一台主机,以保证 PHDA 门户网站和检索系统正常对外提供服务。

7.4 成员责任

数据访问工作人员的职责包括:

(1) 数据服务管理员负责跟踪用户需求,解答用户疑问,处理用户订单,向用户提供数据产品。

(2) 运维工程师负责维护发布数据库、大文件传输工具、PHDA 网站等信息系统和应用软件,确保系统安全稳定运行。

8 系统管理

8.1 管理功能

系统管理为 PHDA 的整体运行提供服务,包括系统管理配置、访问控制、用户行为监控、摄入策略制定、系统运行监控、日志记录等功能,实现数据保存生命周期管理。

PHDA 系统由硬件和软件组成，其中，硬件包括服务器、存储、网络、防火墙等基础设施，软件功能包括数据采集与摄入、数据处理 workflow、数据质量控制、数据管理、数据档案存储和数据访问服务。NPHDC 制定了详细的系统管理配置方案，并部署监控平台对 PHDA 系统硬件设备进行统一监控管理。数据中心使用日志审计系统，全面收集安全设备、网络设备、数据库、服务器、主机等设备所产生的日志（包括运行、告警、操作、消息、状态等），并进行存储、监控、审计、分析、报警、响应和报告。

PHDA 系统实行严格的访问控制，对用户权限进行设置，并采用双因素身份认证，通过 VPN+堡垒机对运维人员进行集中管理。系统依据访问控制策略监控访问行为中的异常情况，如未授权用户操作数据或试图更改系统配置、用户进行恶意下载等，系统记录异常日志并发出告警。系统监控每次用户请求的响应结果和响应时间，如果出现大范围的获取失败或时间延迟，应立即通知管理员。

PHDA 系统记录用户访问、登录、注册、汇交数据、下载数据等操作行为信息，形成详细的日志，以便统计分析。其中摄入行为相关信息包括操作人、时间、摄入条目、成功/失败条目、备份位置等，并发送摄入报告；用户访问行为的信息包括用户名、时间、行为、获取的信息、是否成功、IP 地址等。

8.2 管理活动

PHDA 系统的操作、维护和演进包括以下管理活动：

（1）PHDA 技术和管理团队(数据存档管理员、数据摄入管理员、运维工程师和研究人员等)每周召开例会讨论数据集管理问题。

（2）NPHDC 召集所有工作人员，定期召开数据中心内部跟进会议。

（3）NPHDC 的工作人员与数据生产者保持联系。

（4）每年召开理事会，为 PHDA 建设提供高水平的指导方针和建议。

（5）NPHDC 提供联系电话（010-65250541）和邮箱（rkjkpt@126.com），以及技术支持电话（010-52328766），帮助用户使用 PHDA 服务，并提供关于内容和访问故障或改进建议的反馈。NPHDC 负责报告并在短时间内答复来电者。用户在数据提交过程中可以联系专门的数据摄入管理员。

8.3 成员责任

NPHDC 为 PHDA 的运行和维护设置了不同的角色，包括数据摄入管理员、数据存档管理员、数据服务管理员、同行专家、运维工程师等。每个角色需要根据相应的工作流程 and 操作规范进行运维活动，其中操作人员需要完成以下活动：

（1）操作过程中发现异常事件应及时通知有关部门，并按操作规程和故障计划进行处理。

（2）按操作计划进行监测和日常维护。

（3）记录运行和维护活动的事件。

（4）收集和归档运行维护文件。

（5）定期进行分析和评价，并提出优化和改进建议。

9 保存规划

NPHDC 计划对数据档案制定保存和长期保存计划，采取一系列技术措施进行数据管理和信息系统管理，并对这些措施保持更新，以防止过时风险，保障数据长期可用。

9.1 数据可持续性

NPHDC 实现了基于数据模型的数据管理。在数据文件和元数据方面，NPHDC 跟踪行业或领域标准的变化，在必要情况下对数据文件和元数据进行版本升级，以确保数据的可持续使用。

(1) 元数据管理

NPHDC 制定元数据管理策略，除了保留数字对象本身之外，还收集、丰富和处理各种形式的元数据，包括描述性、技术性、管理性、结构性、权限管理、过程相关性等元数据。元数据文件以 XML 格式存储，当元数据标准发生变化时，升级和转换非常容易。这种方法保证了相关数据的长期可解释性、可检索性和互操作性。

(2) 格式管理

数据资源存储格式的可持续性，直接关系到资源的真实性、完整性与可靠性，是数据资源存储和长期保存中最关键的问题。数据中心通过规范化地开展格式识别、验证与风险评估，来保障格式可持续性。

A.格式识别是指对数字档案资源的技术属性，包括生成环境等进行识别和描述，一般以技术元数据的形式存在。NPHDC 在接收数据资源时即对其格式进行识别，提取技术元数据并保存。

B.格式验证，即通过验证格式现有特征是否符合该格式应有特征来验证其有效性。NPHDC 在接收数据资源时，一旦识别出资源格式，就要对其进行有效性验证。

C.风险评估是指对格式过时风险进行分析，包括格式停产或格式无法读取的风险。NPHDC 将对所有风险格式的数据资源进行规范化格式转化，将信息迁移到可持续的新格式。NPHDC 支持多种数据格式，为了保障数据的长期可访问性、可读性和可用性，列出数据长期保存的推荐格式，如表 2 所示。

表 2 推荐的长期保存文件格式

数据类型	适用于长期保存的文件格式	常用的标准文件格式	来源和应用示例
------	--------------	-----------	---------

音频	WAVE (*.wav), AIFF (*.aiff, *.aif), MXF (*.mxf), Matroska (*.mka)	WAVE (*.wav), MP3 (*.mp3), AAC (*.aac, *.m4a, mp4), AIFF (*.aiff, *.aif), OGG (*.ogg), BWF (*.bwf), MXF (*.mxf), Matroska (*.mka), FLAC (*.flac), OPUS (*.opus)	访谈, 调查
生物材料数据	CSV (*.csv), TXT (*.txt), XML(*.xml)	CSV (*.csv), TXT (*.txt), FASTA (*.fasta), FASTQ (*.fq, *.fastq), PDB(*.pdb, *.ent, *.brk), XLS (*.xls), XML (*.xml)	DNA 测序仪, 质谱仪, 微阵列, 分光光度计
分类, 叙词表, 代码	PDF/A (*.pdf), XML (*.xml)	DOC (*.doc, *.docx), PDF (*.pdf), XML (*.xml)	医学词表、本体
数据库	SQL (*.sql)	SQL (*.sql), SIARD (*.siard), CSV (*.csv), MS Access (*.mdb, *.accdb), dBase (*.dbf), HDF5 (*.hdf5, *.he5, *.h5)	生物医学数据库
地理数据	GML (*.gml), MIF/MID (*.mif/*.mid)	GML (*.gml), MIF/MID (*.mif/*.mid), ESRI Shapefiles (*.shp), MapInfo (*.tab), KML (*.kml)	矢量和栅格数据
影像数据	TIFF (*.tif, *.tiff), PNG (*.png), JPEG2000 (*.jp2), SVG (*.svg)	TIFF (*.tif, *.tiff), JPG (*.jpg, *.jpeg), PNG (*.png), GIF (*.gif), JPEG 2000 (*.jp2), STL (*.stl), OPG (orthopantomogram), DICOM (*.dcm), DVT (digital volume tomography), PDF (*.pdf), SVG (*.svg), Illustrator (*.ai), EPS (*.eps)	来自 MRT 和 CT 扫描, 照相机, 显微镜, 超声仪器以及 X 射线和超声检查设备等成像技术的图像数据
影像数据 3D	OBJ (*.obj, *.mod in ASCII format), X3D (*.x3d), VRML (*.vrml, *.wrl)	STL (*.stl), OBJ (*.obj, *.mod), PLY (*.ply), DXF (*.dxf), X3D (*.x3d), COLLADA (*.dae), FBX (*.fbx), VRML (*.vrml, *.wrl)	立体光刻等 3D 技术

标记语言	XML (*.xml)	XML (*.xml), HTML (*.html), SGML (*.sgml)	元数据
传感器数据	CSV (*.csv), TXT (*.txt), PDF (*.pdf)	CSV (*.csv), TXT (*.txt), PDF (*.pdf), XLS (*.xls, *.xlsx), XML (*.xml)	心电图, 脑电图, 多导睡眠图, 测量血压、血糖水平的设备等。
电子表格	CSV (*.csv)	ODS (*.ods, *.odt, *.odg, *.odc, *.odf), CSV (*.csv), XLS (*.xls, *.xlsx), OOXML (*.docx, *.docm), PDF/A (*.pdf)	来自研究和临床护理的数据
统计数据	R (*.r), CSV (*.csv)	CSV (*.csv), SAS (*.7dat, *.sd2, *.tpst), SPSS (*.sav), SPSS Portable (*.por), STATA (*.dta), DDI (*.xml), data (*.csv, *.txt), R (*.r)	来自临床研究和护理的数据
文本文件	PDF/A (*.pdf), XML (*.xml), Unicode TXT (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py etc. in ASCII format)	DOC (*.doc, *.docx), PDF (*.pdf), TXT (*.txt), RTF (*.rtf), ODT (*.odt), Powerpoint (*.ppt)	医生/医务人员的检查结果报告, 行政数据
视频	Matroska (*.mka), MXF (*.mxf), AVI (*.avi)	MPEG-4 (*.mp4, *.m4a, *.m4v), MPEG-2 (*.mpg, *.mpeg, *.m2v, *.mpg2), AVI (*.avi), MXF (*.mxf), Matroska (*.mka, *.mkv), QuickTime (*.mov, *.qt), Windows Media (*.wmv)	来自影像技术 (例如 CT 扫描, 照相机, 超声波仪器) 的视频文件

9.2 数据完整性

NPHDC 通过校验、冗余存储和及时更换过时的存储介质等措施, 来保障数据完整性。主要包括以下措施:

- (1) 在接收数据资源前首先对其进行扫描, 确保没有恶意软件夹带移交, 并对数据资源进行修复检查, 验证接收状态是否与原始状态一致或是否经过改动。
- (2) 对所有保存的数据资源进行编目, 并记录所有资源的接收过程以及入库后其他生命周期过程, 包括格式迁移、复制以及审计等长期保存过程。
- (3) 对数据资源, 尤其是那些存储载体有过时或损毁风险的资源进行复制

备份,必要的情况下可以使用阻滞剂来防止物理载体损毁对数字内容造成的破坏。

(4) 对保存的数据资源进行年度抽检,包括载体抽检和修复检查,对抽检有问题的资源进行载体更换或修复。

(5) 所有存储载体保存至第十年,必须将其中的数字内容保存到新的可靠载体上。

(6) 对 PHDA 系统进行季度日志审查,验证数据资源内容、结构、背景信息的完整性,及其管理过程可追溯。

9.3 服务连续性

PHDA 需要跟踪信息技术的发展和数据用户需求的变化,并保持服务可持续性。NPHDC 通过在线用户反馈机制、领域专家咨询和 IT 技术专家咨询,收集系统更新需求。系统升级通常每 3-5 年进行一次。PHDA 的升级将由专业软件公司按照软件工程管理要求进行。在 PHDA 新版本发布之前,NPHDC 将开展多轮系统测试,以确保系统安全性和可靠性。测试通过后,NPHDC 将根据迁移计划实现系统新版本部署和数据迁移,确保数据和相关信息的完整性和一致性。

(1) 资金保障及所属机构的支持

国家人口健康科学数据中心(NPHDC)是国家科技部和财政部认定的 20 个国家科学数据中心之一,属于国家科技基础条件平台下的科技资源共享服务平台,依托单位是中国医学科学院。NPHDC 建设经费由科技部统筹、财政部拨款,每年可获得科技部的固定运营资金。NPHDC 的工作人员大多为中国医学科学院的在编员工,可以有效保证工作人员的稳定性。此外,中国医学科学院还为 NPHDC 提供 A 级机房和固定的工作场所。上述支撑条件可以确保 NPHDC 持续提供数据访问服务。

(2) 跟踪和评估数据管理标准和数据规范的发展趋势

NPHDC 建立动态监测系统,跟踪国内外前沿数据政策、标准和管理需求的发展趋势和新要求。这些跟踪结果为 PHDA 相关标准(如 AIP、SIP 要求)和数据政策提供了重要参考。NPHDC 跟踪清单如下:

A.人口健康科学数据共享数据集分类与编码标准

B.常见的数据格式

C.GB/T 32843-2016 科技资源鉴定

D.GB/T 36344-2018 信息技术数据质量评价指标

E.GB/T 30523 科技资源核心元数据

F.GB/T 31073 科技资源服务核心元数据

.....

(3) 跟踪和评估技术系统的发展趋势

NPHDC 系统建设遵循以下国家标准和国际标准。NSSDC 对以下标准的制定进行长期跟踪,并根据跟踪结果对技术系统升级需求进行评估。

A.ISO 14721 开放档案信息系统参考模型

B.ISO 16363 可信数字仓储的审计及认证：指标与列表
C. ISO 20652-2006 生产者-档案接口——方法抽象标准
D.GB/T 25058-2019 信息安全技术 网络安全等级保护实施指南
E.GB/T 22239-2019 信息安全技术——网络安全分类保护基准
F.GB 4943.1-2011 信息技术 设备安全 第 1 部分：通用要求
G.GB 4943.23-2012 信息技术 设备安全 第 23 部分：大数据存储设备
H.GJB9001C 质量管理体系要求

.....

(4) 跟踪用户服务需求和产品技术

为了改进服务，NPHDC 采用如下方法来跟踪用户(生产者和消费者)的服务需求和产品技术。

A.在数据存档计划准备期间跟踪需求。在数据存档计划的准备过程中，NPHDC 的存档管理员与数据生产者就存档数据格式、版本、专用软件工具、存档安排、存档接口等进行协商。存档管理员在每次存档过程中确认这些内容。该过程是 NPHDC 跟踪数据生产者服务和产品技术需求的有效途径之一。

B.用户反馈和调查活动。NPHDC 通过学术会议、网站反馈和用户调查活动跟踪用户(生产者和消费者)的需求。