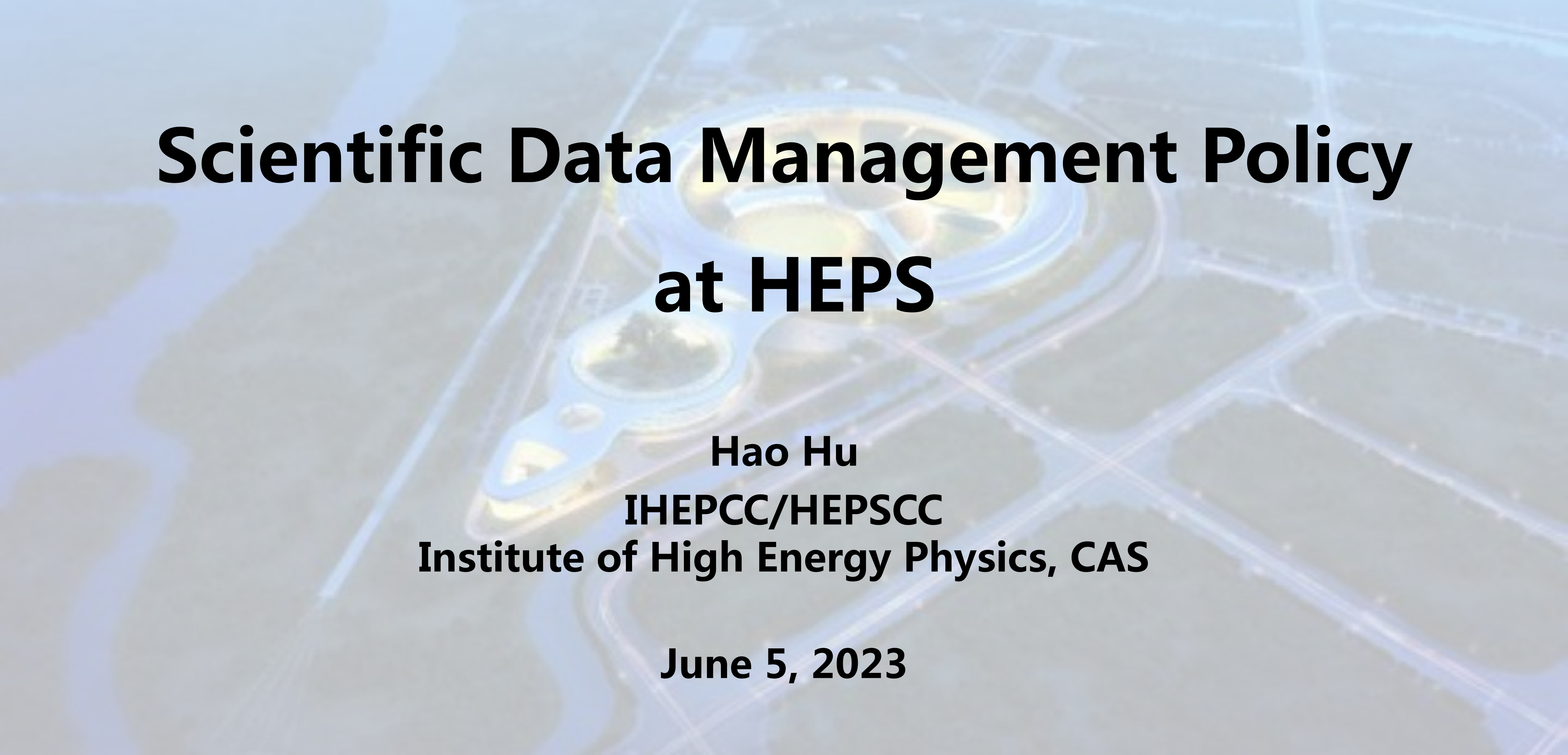
中国科学院高能物理研究所

Institute of High Energy Physics, Chinese Academy of Sciences



Scientific Data Management Policy at HEPS

Hao Hu

IHEPCC/HEPSCC

Institute of High Energy Physics, CAS

June 5, 2023

Outline

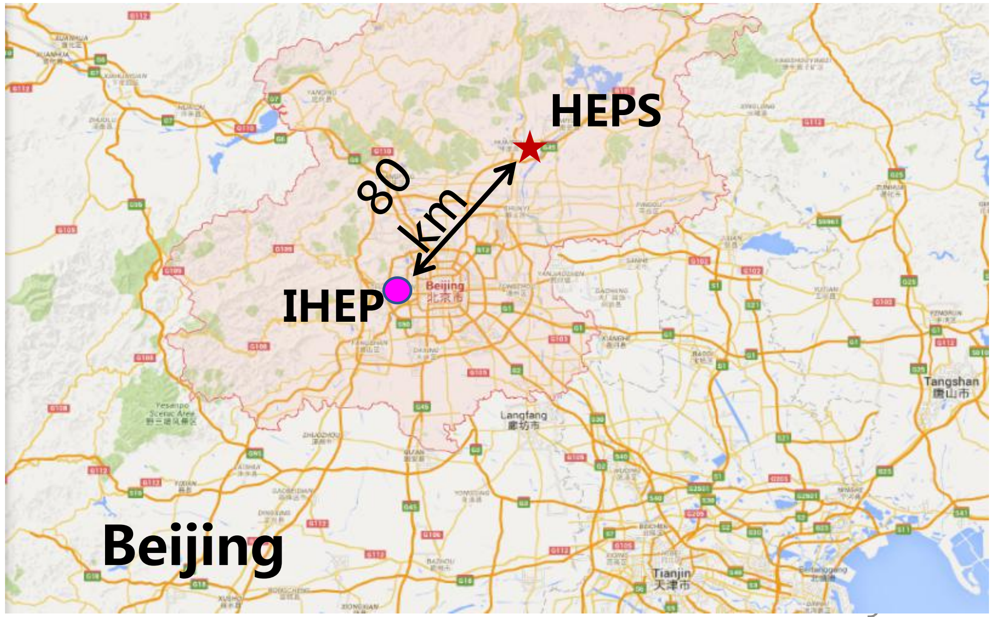
1. HEPS Introduction
2. Data challenge and the motivation
3. Overview of the data policy framework
4. The specific details of the data policy
5. Consideration & Discussion

High Energy Photon Source (HEPS)

中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

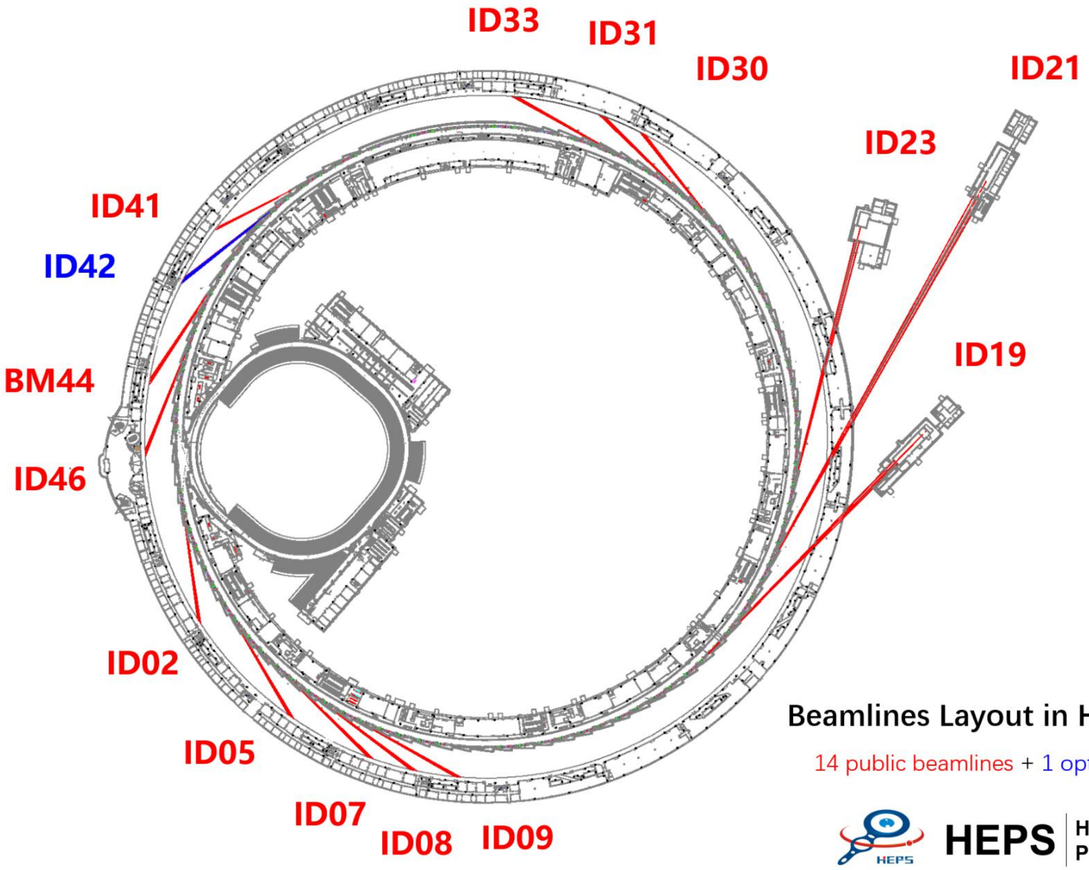
- The fourth generation light source in China — High energy, high brightness
- Located in Beijing - about 80KM from IHEP
- Officially approved in Dec. 2017
- The construction was started at the end of 2018
- The whole project will be finished in mid-2025

Main parameters	Unit	Value
Beam energy	GeV	6
Circumference	m	1360.4
Emittance	pm·rad	< 60
Brightness	phs/s/mm^2/mrad^2/0.1%BW	>1x10^{22}
Beam current	mA	200
Injection		Top-up



HEPS Beamlines in phase I

中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences



Beamlines Layout in HEPS phase I

14 public beamlines + 1 optics test beamline

HEPS HIGH ENERGY PHOTON SOURCE

Microfocusing X-Ray Protein Crystallography-ID02 Beamline

Low-Dimensional Structure Probe Beamline-ID05

Engineering Materials Beamline-ID07

Hard X-Ray Coherent Scattering Beamline-ID09

Pink Beam SAXS Beamline-ID08

Hard X-Ray Nanoprobe Multimodal Imaging-ID19 Beamline

Hard X-Ray Imaging Beamline-ID21

Structural Dynamics Beamline-ID23

ID30-Transmission X-Ray Microscopic Beamline

ID31-High Pressure Beamline

ID33-Hard X-Ray High Resolution Spectroscopy Beamline

BM44-Tender X-Ray Beamline

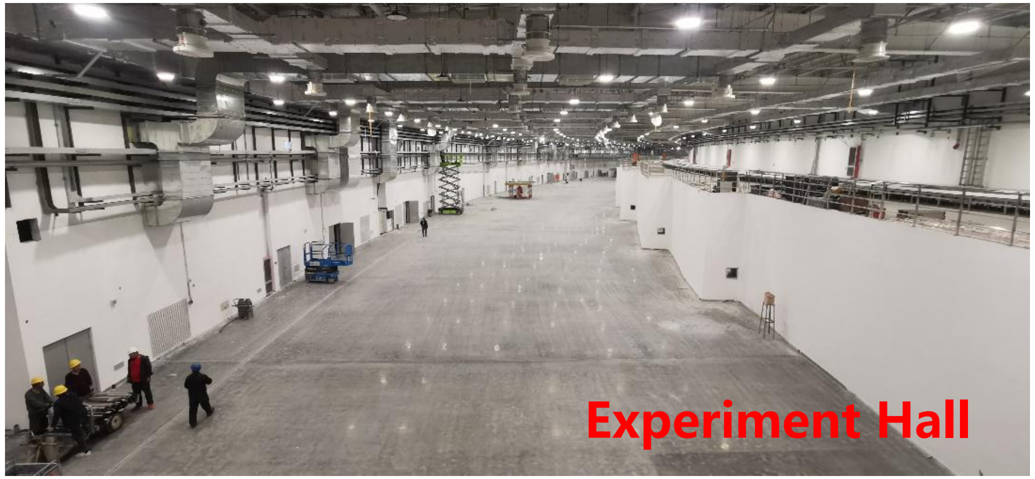
ID41-High Resolution Nanoscale Electronic Structure Spectroscopy Beamline

ID42-Optics Test Beamline

ID46-X-Ray Absorption Spectroscopy Beamline

14 public beamlines + 1 optics test beamline in Phase I

Can accommodate over 90 beamlines in total


Experiment Hall

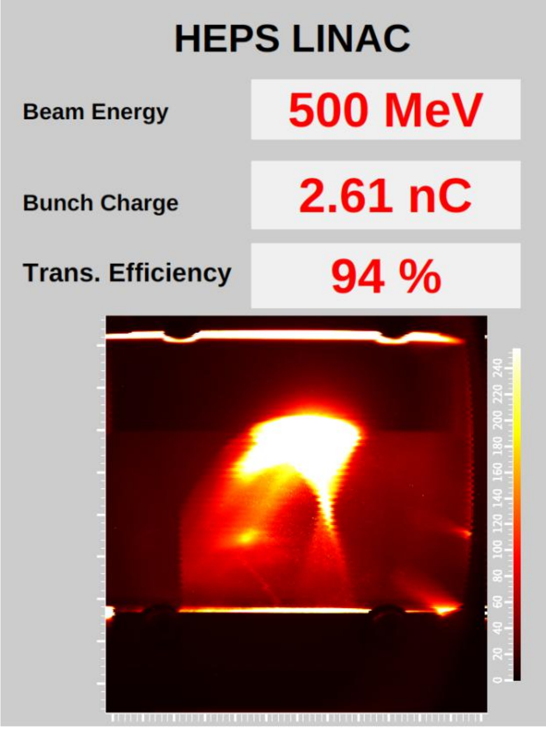
Progress of the HEPS project

□ The construction of the civil structure completed. Now at the stage of equipment installation

□ 2023.01, HEPS booster installation completed

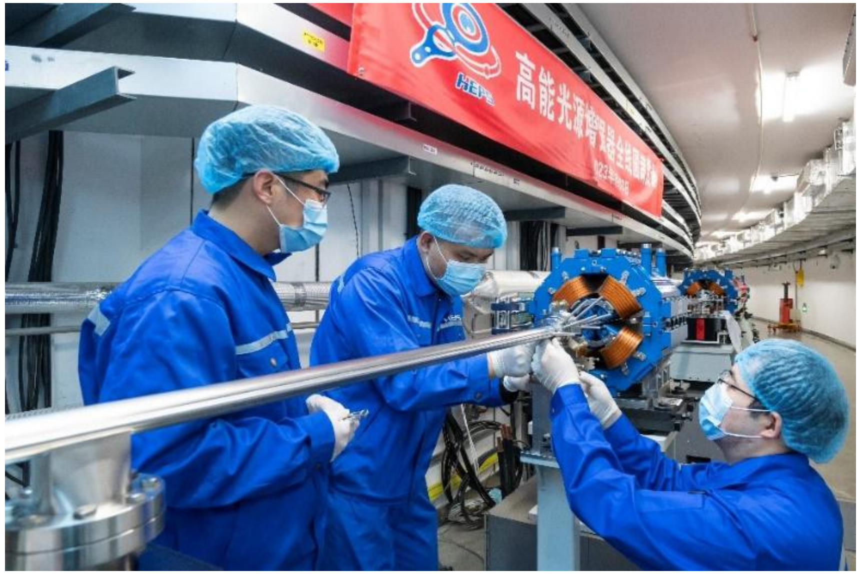
□ 2023.02, Start installation of storage ring

□ 2023.03, HEPS achieved the first electron beam accelerated to 500 MeV



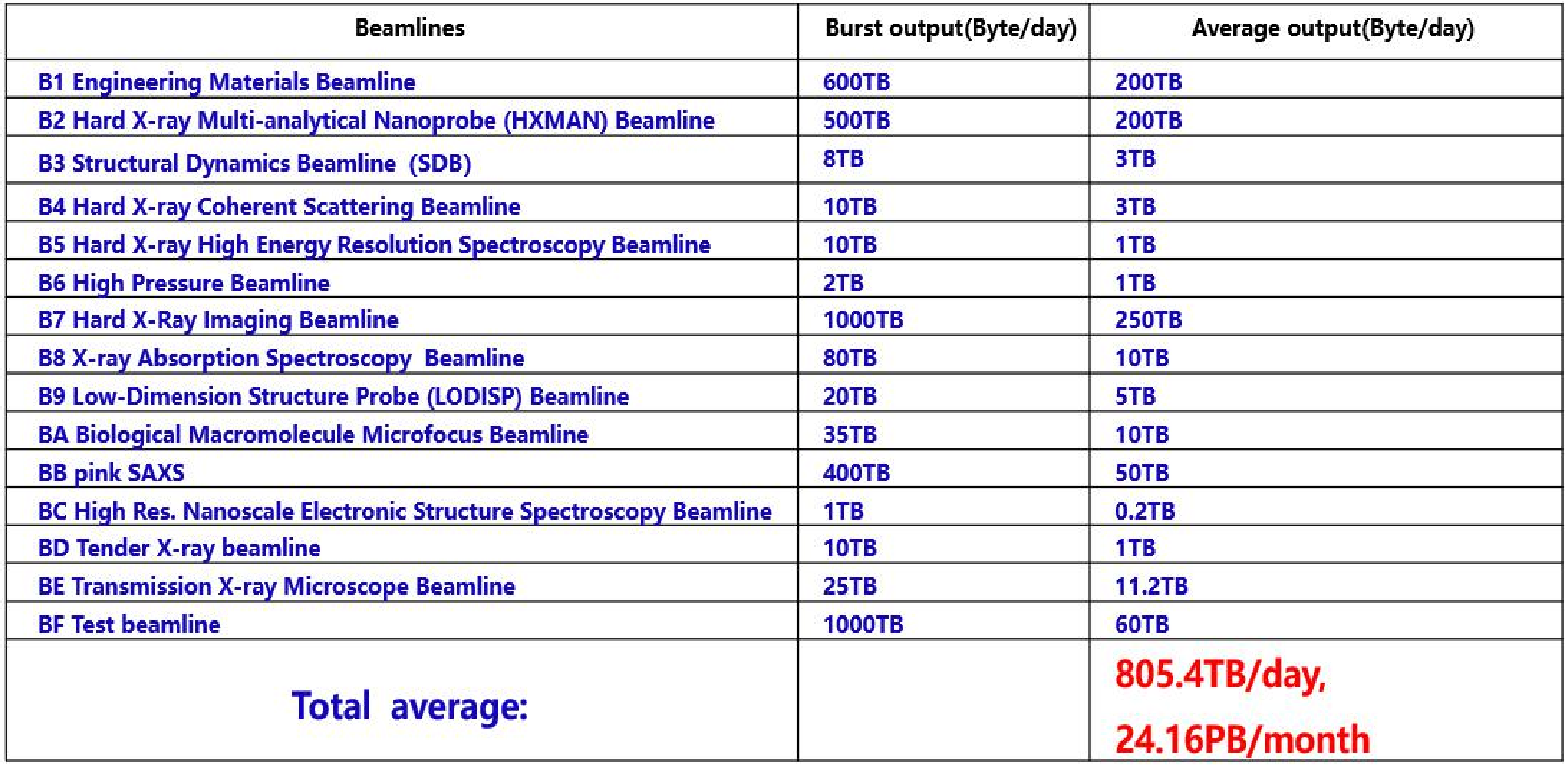
HEPS LINAC

Beam Energy	500 MeV
Bunch Charge	2.61 nC
Trans. Efficiency	94 %



Data Challenges @HEPS

- □ Increased source brightness, X-ray detector capabilities have been continuously improving
- □ **More than 24PB raw data will produced per month**



Beamlines	Burst output(Byte/day)	Average output(Byte/day)
B1 Engineering Materials Beamline	600TB	200TB
B2 Hard X-ray Multi-analytical Nanoprobe (HXMAN) Beamline	500TB	200TB
B3 Structural Dynamics Beamline (SDB)	8TB	3TB
B4 Hard X-ray Coherent Scattering Beamline	10TB	3TB
B5 Hard X-ray High Energy Resolution Spectroscopy Beamline	10TB	1TB
B6 High Pressure Beamline	2TB	1TB
B7 Hard X-Ray Imaging Beamline	1000TB	250TB
B8 X-ray Absorption Spectroscopy Beamline	80TB	10TB
B9 Low-Dimension Structure Probe (LODISP) Beamline	20TB	5TB
BA Biological Macromolecule Microfocus Beamline	35TB	10TB
BB pink SAXS	400TB	50TB
BC High Res. Nanoscale Electronic Structure Spectroscopy Beamline	1TB	0.2TB
BD Tender X-ray beamline	10TB	1TB
BE Transmission X-ray Microscope Beamline	25TB	11.2TB
BF Test beamline	1000TB	60TB
Total average:		805.4TB/day, 24.16PB/month

Estimated data volume of HEPS at Phase I

Huge amount of data is a big challenge for data management and processing

Why need data policy for HEPS?

□ To address massive data challenge

- Facility will provide the service for data curation and access
- Follow FAIR Principles (Findability, Accessibility, Interoperability, and Reuse)
- Clarify the responsibilities and obligations of facility and users

□ National data policy requirement

- In 2018, The Chinese government issued the “MEASURES OF SCIENCE DATA MANAGEMENT”
- In 2019, the Chinese Academy of Science issued the "Measures for the Management and Open Sharing of Scientific Data in CAS(Trial)"

□ Current Status

- No practices and regulations about data openness and sharing are available for user facilities so far in China
- No sufficient policy framework to support data management for HEPS

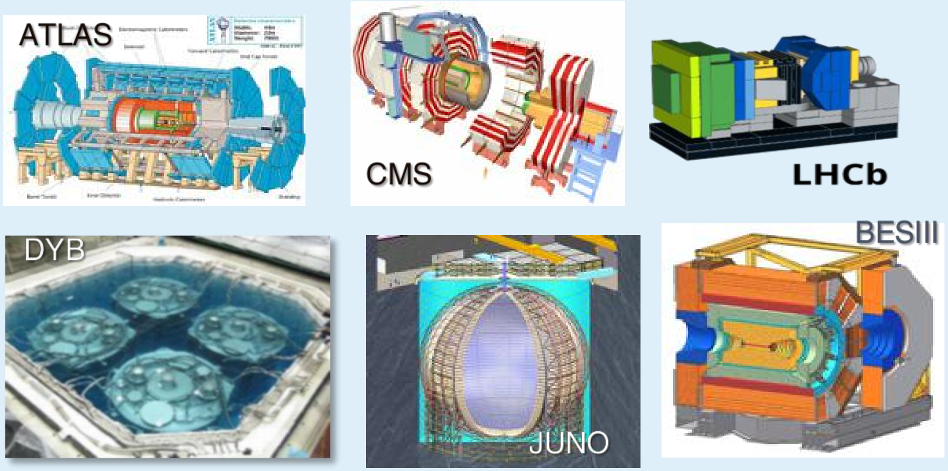
□ To develop data policy for HEPS

□ As guidelines for the design and implementation of data management

□ To convince users that their intellectual property rights is protected

Data policy for High Energy Physics Experiment

- □ Members of HEP Projects come from international collaboration groups
- □ Scientific data are shared among collaboration group members
- □ Data board select a part of raw data and result data to be openly accessible

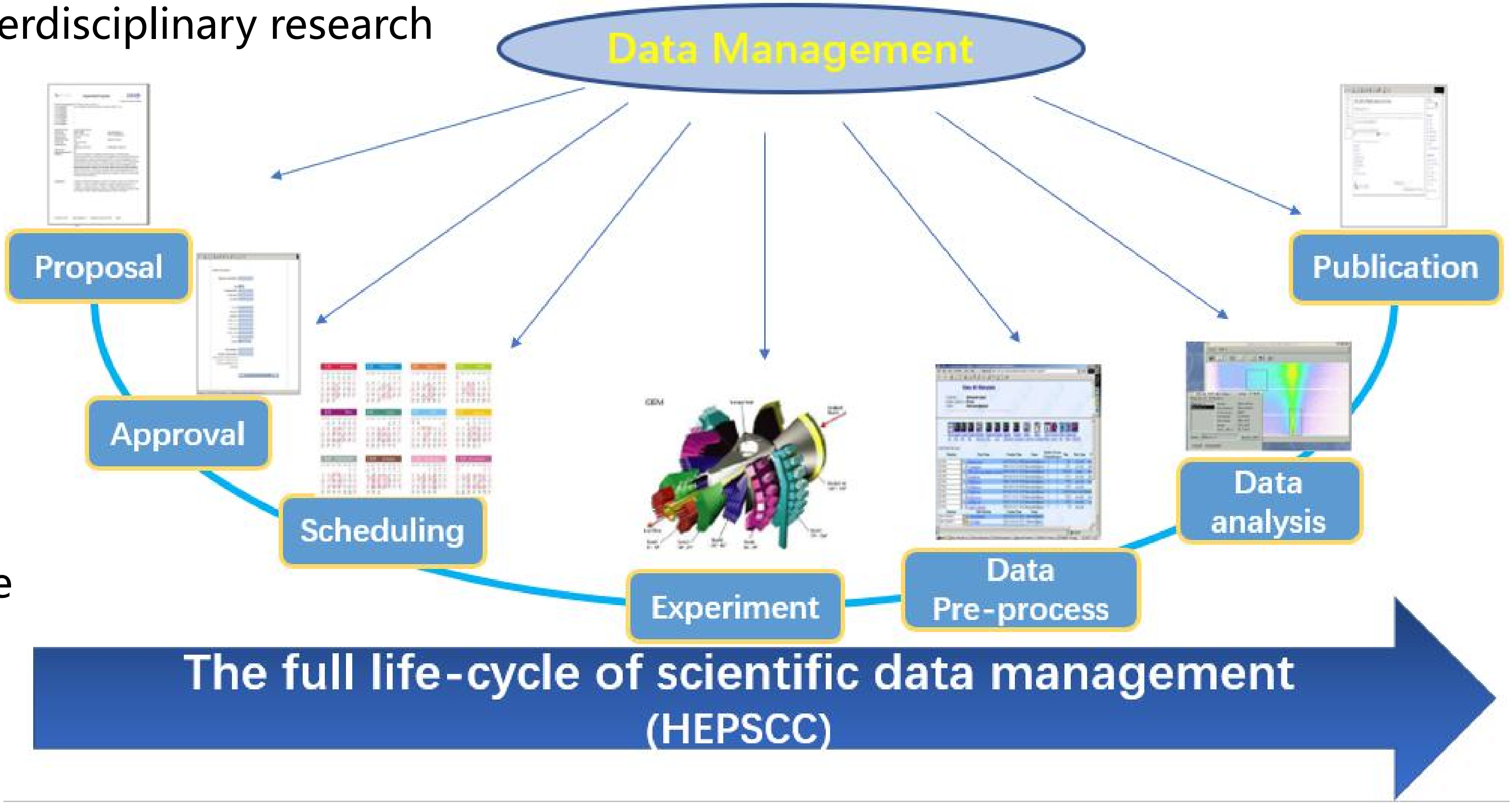


HEP Experiments hosted and participated by IHEP

Scientific data lifecycle at HEPS

HEPS is a public experimental platform for interdisciplinary research

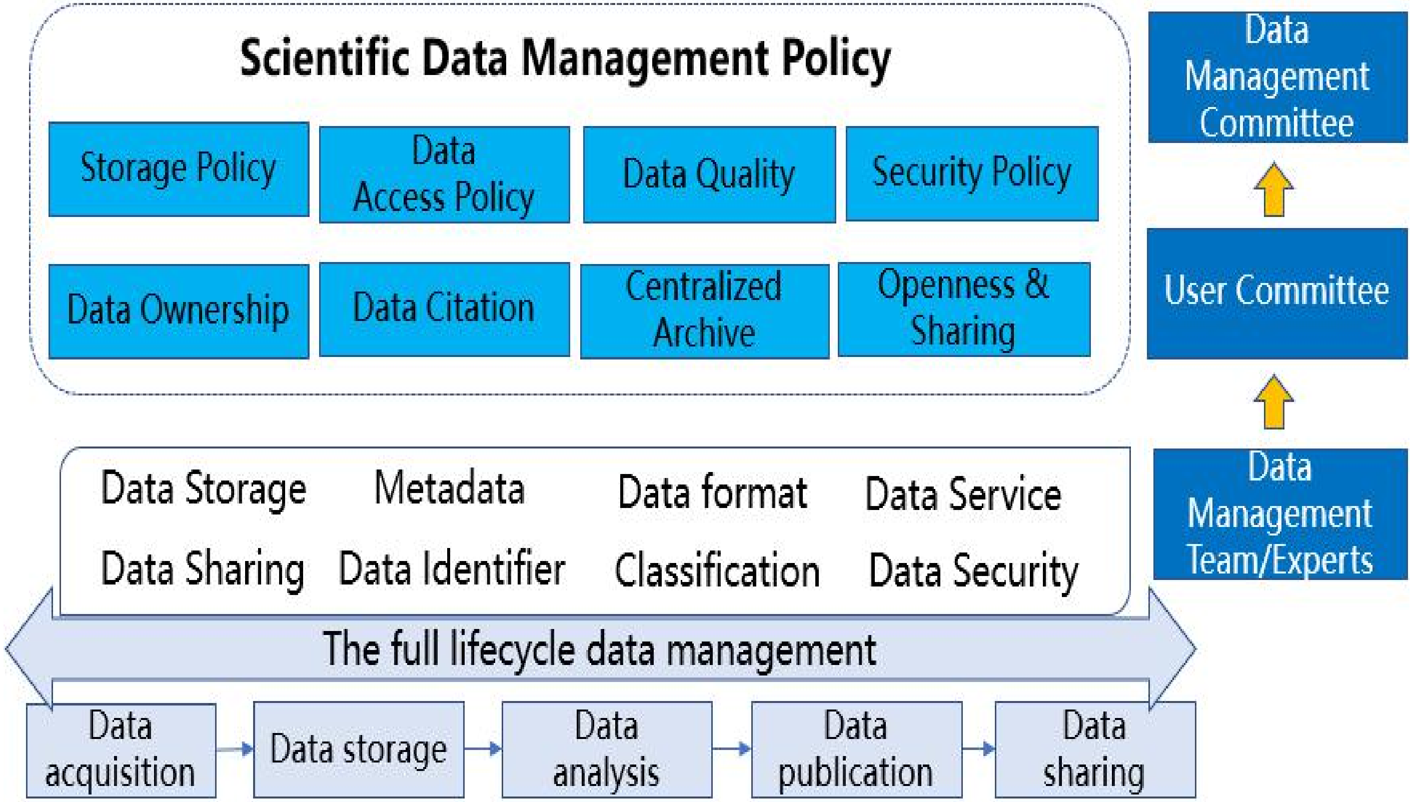
1) The principal investigator (PI) submits a proposal
2) After the peer review, the proposal is approved by the management board
3) HEPS allocates a beamtime to the PI
4) The PI' s team do the experiment at a beamline station
5) Raw data produced from the experiment are saved to the HEPS storage
6) During/after the experiment, users can use the computing resources provided by HEPS for data analysis
7) When data are analyzed, the processed data are also saved to HEPS storage
8) The PI will publish their scientific discovery and associated data
9) After the embargo period, the data will be opened and shared with other researchers



HEPSCC(HEPS Computing & Communication)

Members come from IHEP, responsible for the data curation and provide access to scientific data of HEPS

Data Policy Framework



- □ **Develop standards and policies around the full lifecycle of data management**
- □ **DM tasks: data classification, storage, data format, data service, data sharing, data PID, data security…**
- □ **Data policy includes policies about data storage, data access, data service guarantee, openness and sharing…**
- □ **Decision-making process**

  DM team/Experts → User Committee → Data Management Committee

General principles

- About the ownership, curation, archiving and access to scientific data and metadata
- Acceptance of this Data Policy is a condition for the award of beamtime
- Users must not attempt to access, exploit or distribute scientific data or metadata illegally without authorization
- HEPS reserves the right to deny access to scientific data and any future beamtime from a user who violates the Data Policy
- Data classification: raw data, processed data, calibration data, result data…
- HEPS Facility users must be officially registered
- HEPS is obligated to keep user details secure
- Proprietary research is not covered by this data policy

Storage policy for scientific data

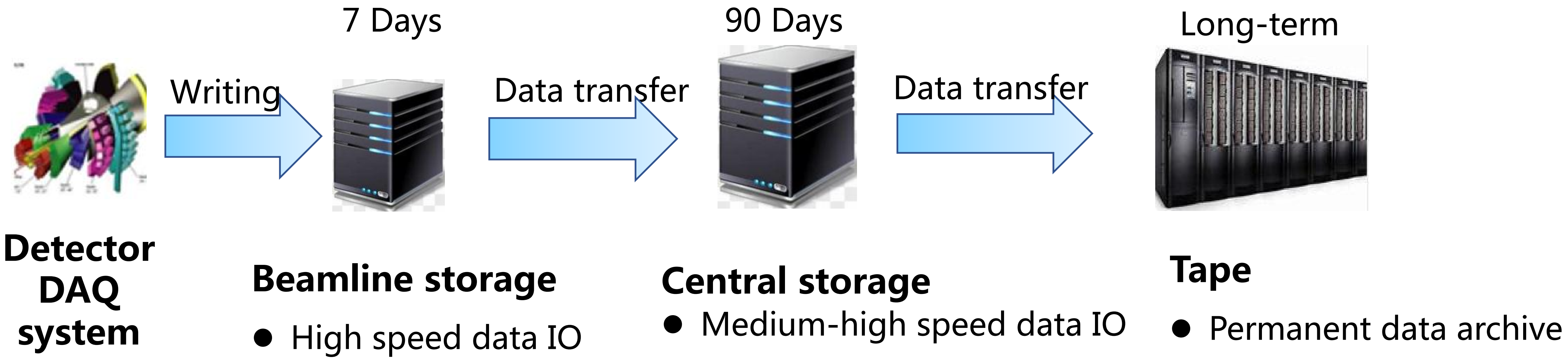
HEPS supports hierarchical storage architecture

① **Raw data/processed data are kept on beamline storage for up to 7 days**

② **Raw data/processed data are kept on central storage for up to 90 days**

③ **Raw data are archived to the tape for Long-term storage**

Data storage policy will be adjusted according to the actual data volume and funding situation



Detector DAQ system

Beamline storage

- High speed data IO

Central storage

- Medium-high speed data IO

Tape

- Permanent data archive

Curation and Access for Raw data

- **Data Format**
  - Well-defined format(HDF5)
  - HEPSCC provide tools to transform to other data format
- **Raw data are read-only during the storage**
- **Metadata will be curated in raw data files and metadata catalogue**
- **Raw data can be accessed by searching metadata catalogue**
- **Each dataset will have a unique persistent identifier**
  - CSTR(Chinese Science and Technology Resource)
  - DOI(Digital Object Identifier)
- **Embargo period**
  - 2 years by default
  - PI can request to extend it

Curation and Access for processed data and results

□ **Ownership**

- Determined by the person performing the analysis

□ **Storage**

- Processed data and results data will not be long-term preserved
- Calibration and alignment data will be long-term preserved

□ **Access**

- Result data are restricted to experiment team, unless requested by PI
- Calibration and alignment data can be openly accessible, not restricted by embargo period

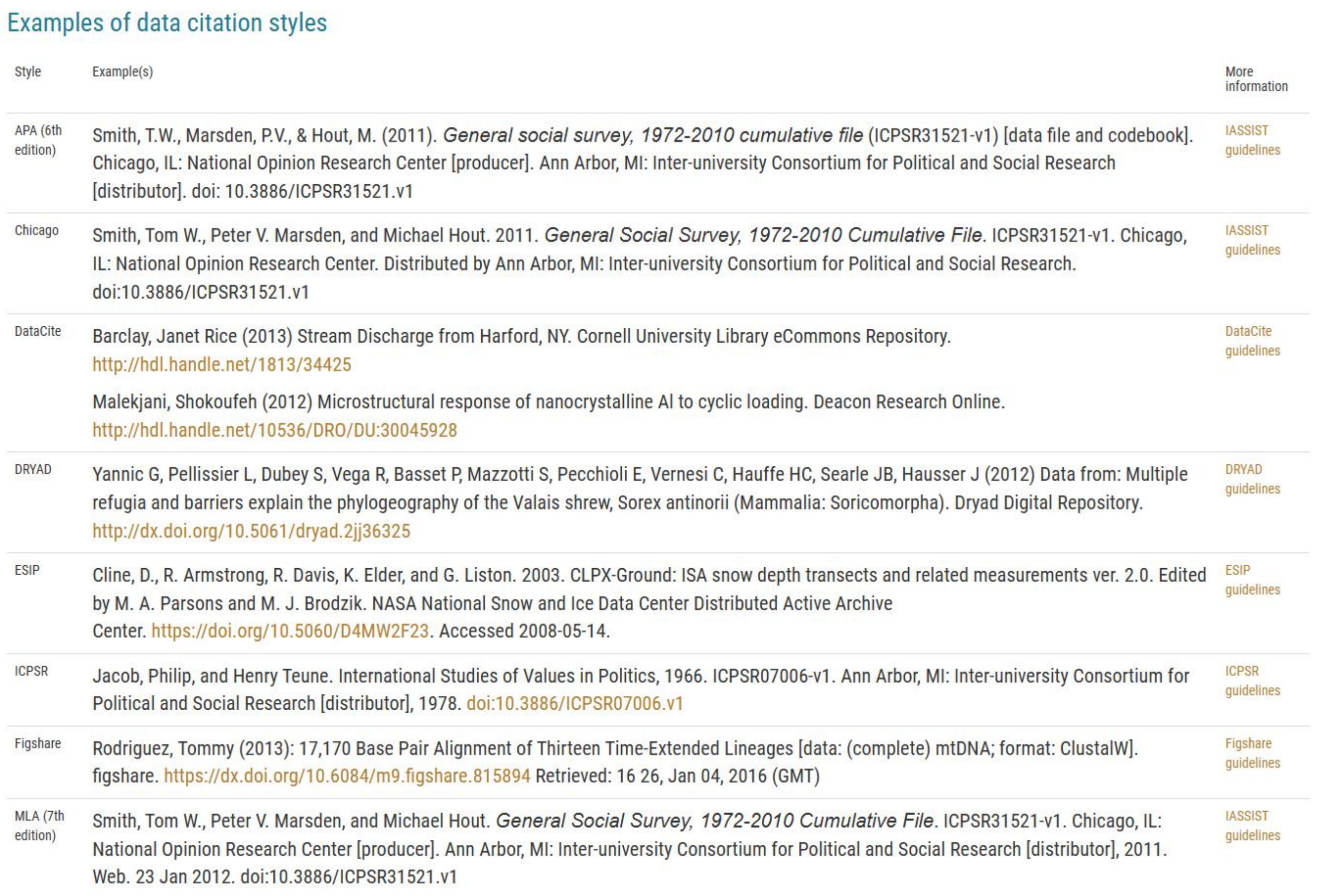
Publication and Citation

□ Publication

- Publications references related to experiments carried out at HEPS should be deposited in the publications database

□ Citation

- Any publication use the data should cite the persistent ID of the data



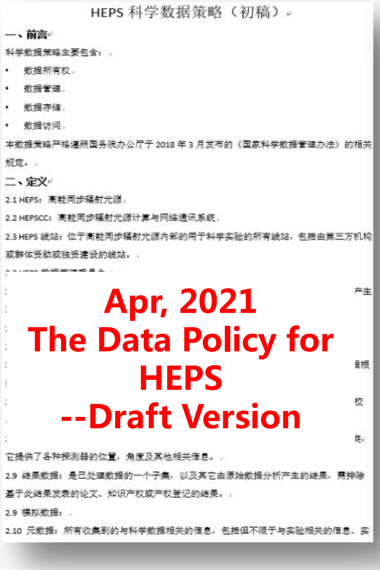
Examples of data citation styles

Style	Example(s)	More information
APA (6th edition)	Smith, T.W., Marsden, P.V., & Hout, M. (2011). *General social survey, 1972-2010 cumulative file* (ICPSR31521-v1) [data file and codebook]. Chicago, IL: National Opinion Research Center [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. doi: 10.3886/ICPSR31521.v1	IASSIST guidelines
Chicago	Smith, Tom W., Peter V. Marsden, and Michael Hout. 2011. *General Social Survey, 1972-2010 Cumulative File*. ICPSR31521-v1. Chicago, IL: National Opinion Research Center. Distributed by Ann Arbor, MI: Inter-university Consortium for Political and Social Research. doi:10.3886/ICPSR31521.v1	IASSIST guidelines
DataCite	Barclay, Janet Rice (2013) Stream Discharge from Harford, NY. Cornell University Library eCommons Repository. http://hdl.handle.net/1813/34425 Malekjani, Shokoufeh (2012) Microstructural response of nanocrystalline Al to cyclic loading. Deacon Research Online. http://hdl.handle.net/10536/DRO/DU:30045928	DataCite guidelines
DRYAD	Yannic G, Pellissier L, Dubey S, Vega R, Basset P, Mazzotti S, Pecchioli E, Vernesi C, Hauffe HC, Searle JB, Hausser J (2012) Data from: Multiple refugia and barriers explain the phylogeography of the Valais shrew, Sorex antinorii (Mammalia: Soricomorpha). Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.2jj36325	DRYAD guidelines
ESIP	Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. A. Parsons and M. J. Brodzik. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5060/D4MW2F23. Accessed 2008-05-14.	ESIP guidelines
ICPSR	Jacob, Philip, and Henry Teune. International Studies of Values in Politics, 1966. ICPSR07006-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1978. doi:10.3886/ICPSR07006.v1	ICPSR guidelines
Figshare	Rodriguez, Tommy (2013): 17,170 Base Pair Alignment of Thirteen Time-Extended Lineages [data: (complete) mtDNA; format: ClustalW]. figshare. https://dx.doi.org/10.6084/m9.figshare.815894 Retrieved: 16 26, Jan 04, 2016 (GMT)	Figshare guidelines
MLA (7th edition)	Smith, Tom W., Peter V. Marsden, and Michael Hout. *General Social Survey, 1972-2010 Cumulative File*. ICPSR31521-v1. Chicago, IL: National Opinion Research Center [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011. Web. 23 Jan 2012. doi:10.3886/ICPSR31521.v1	IASSIST guidelines

Examples of Data citation styles

Progress

- □ In Apr 2021, The Data Policy for HEPS --Draft Version was finished
- □ In Mar 2023, we have a revised draft version
- □ Need to be approved by HEPS Data Management Committee
- □ Data policy is the guideline for the data management for HEPS
- □ Data management system supports the policy implementation
  - User Service Portal(user authentication & authorization)
  - Metadata catalogue(storage policy)
  - Data Portal(embargo period, data openness and sharing)
  - Data transfer(storage policy)
  - Data format design



HEPS 科学数据策略（初稿）

一、前言

科学数据策略主要包含：

- 数据所有权
- 数据管理
- 数据存储
- 数据访问

本数据策略严格遵照国务院办公厅于 2018 年 3 月发布的《国家科学数据管理办法》的相关规定。

二、定义

2.1 HEPS：高能同步辐射光源

2.2 HEPSCC：高能同步辐射光源计算与网络通讯系统

2.3 HEPS 线站：位于高能同步辐射光源内部的用于科学实验的所有线站，包括由第三方机构或群体资助或独资建设的线站。

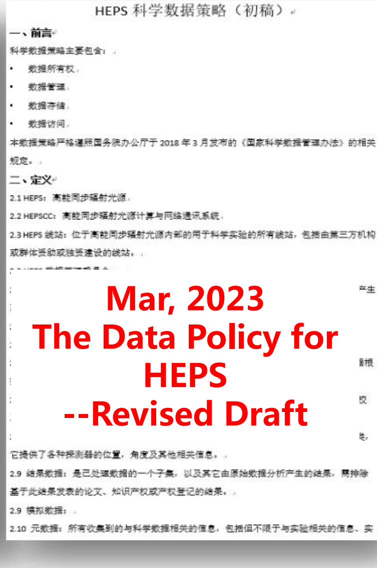
Apr, 2021
The Data Policy for HEPS
--Draft Version

它提供了各种探测器的位置，角度及其他相关信息。

2.9 结果数据：是已处理数据的一个子集，以及其它由原始数据分析产生的结果，需排除基于此结果发表的论文、知识产权或产权登记的结果。

2.9 模拟数据：

2.10 元数据：所有收集到的与科学数据相关的信息，包括但不限于与实验相关的信息、实



HEPS 科学数据策略（初稿）

一、前言

科学数据策略主要包含：

- 数据所有权
- 数据管理
- 数据存储
- 数据访问

本数据策略严格遵照国务院办公厅于 2018 年 3 月发布的《国家科学数据管理办法》的相关规定。

二、定义

2.1 HEPS：高能同步辐射光源

2.2 HEPSCC：高能同步辐射光源计算与网络通讯系统

2.3 HEPS 线站：位于高能同步辐射光源内部的用于科学实验的所有线站，包括由第三方机构或群体资助或独资建设的线站。

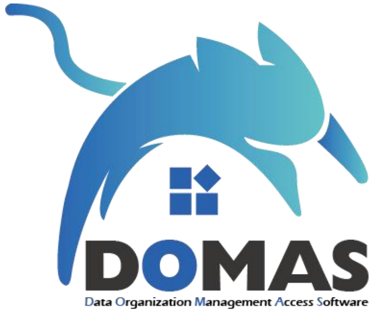
Mar, 2023
The Data Policy for HEPS
--Revised Draft

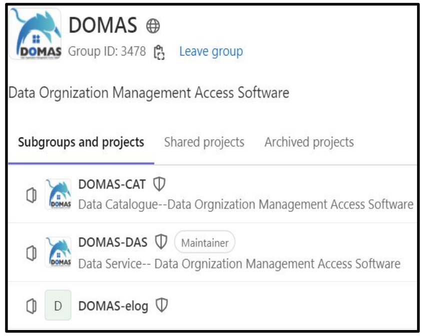
它提供了各种探测器的位置，角度及其他相关信息。

2.9 结果数据：是已处理数据的一个子集，以及其它由原始数据分析产生的结果，需排除基于此结果发表的论文、知识产权或产权登记的结果。

2.9 模拟数据：

2.10 元数据：所有收集到的与科学数据相关的信息，包括但不限于与实验相关的信息、实





DOMAS
Group ID: 3478 Leave group

Data Orgnization Management Access Software

Subgroups and projects | Shared projects | Archived projects

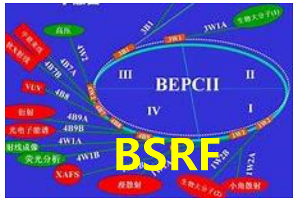
DOMAS-CAT
Data Catalogue--Data Orgnization Management Access Software

DOMAS-DAS Maintainer
Data Service-- Data Orgnization Management Access Software

DOMAS-elog

HEPS Data Management System

Photon/Neutron Source Facility Alliance for data and software



Conference of Advanced Photon/neutron Source Data And Software(CAPSDAS)
Mar, 2023 • BEIJING

Alliance founding members

- HEPS (High Energy Photon Source)
- SHINE (Shanghai HIgh repetitioN rate XFEL and Extreme light facility)
- SSRF (Shanghai Synchrotron Radiation Facility)
- HLS (Hefei Light Source)
- CSNS (China Spallation Neutron Source)

Collaborate to address data and software challenges

- Establish common scientific data management policy
- Develop metadata standard
- R&D of data management and analysis software framework
- Develop disciplinary algorithm and software
- Build software ecosystem

Consideration & Discussion

Discussed with data management experts, beamline scientists, Information specialists, need to be further discussed:

- **Should raw data be long-term preserved? Or be the pre-processed data?**
- **How to ensure the integrity and accuracy of metadata for further utilization?**
  - The PI has the responsibility to ensure the correctness of metadata?
- **Data policy does not have policy about simulated data currently**
- **HEPS Data Management Committee need to be set up as soon as possible**
- **The HEPS budget does not cover long-term storage, but we hope it will be supported by the nation**

Thank you for your attention!

Comments or suggestions?