NDS
National Data Storage

UNIVERSAL INFRASTRUCTURE FOR SHARING, STORAGE AND EFFECTIVE PROCESSING OF LARGE DATA VOLUMES IN HPC, BIG DATA AND ARTIFICIAL INTELLIGENCE MODELS

DATA SCIENCE
central data silos disposal

SHARED ACCESS
one access for all interested parties

USER SUPPORT
expertise on data management and infrastructure to help researchers

BACKGROUND

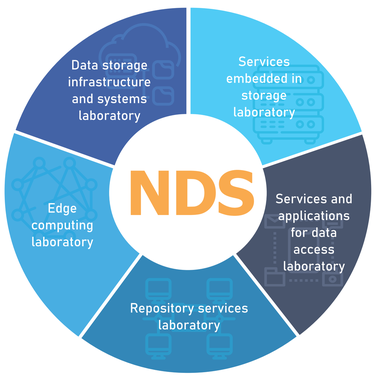
At the heart of the National Data Storage project is a need to provide services and **data infrastructure** for the entire community in Poland: academic users, governmental institutions and local administration. In addition, the NDS reserves its 30% of capacity for industry access. Basic and key for the performance of **R&D** and serving environments in the area of **academics** (universities, Polish Academy of Sciences) and industry. The need to provide commonly available and **utilitarian resources for data storage** was both noticed by the Ministry of Education and Science in Poland and by the European Commission in Europe.

The hope placed in the NDS was confirmed by several **support letters** received from many consortiums from the Polish Map of Research Infrastructures and European Research Infrastructures from the **ESFRI map** (European Strategy Forum for Research Infrastructure), incl. POLFEL, POLFAR, Elixir.pl, UVSat, EURO-ARGO, ECBIG, LSST.pl, BBMRI.pl, as well as the largest

ARCHITECTURE

NDS software platform uses the **Data Lake** concept. The openness of the system is a key feature that distinguishes it from other existing solutions. The implementation of this concept enables t**he storage of extensive data in raw format** and the integration of many services and protocols for data storage and access, as well as additional functions related to the presentation, conversion, discovery and processing of data, integrated as services and applications or environments for **data analysis and processing.**

The NDS system architecture ensures the storage, **protection,** sharing and processing of **scientific and business data**, ensuring **high level of data security** and management processes. It also enables the interoperability and synergistic cooperation of the data infrastructure with specialized external systems, including national HPC, cloud, **BigData and AI infrastructures.**



The versatility and scalability of the architecture of the NDS will allow for the implementation of tasks in this system for both great challenges of science and mass services for individual users.

In addition, **data-centric services** and applications will be integrated with the data storage system, including environments supporting **Data Science** and **group collaboration** based on data, e.g. collaborative editing of documents, diagrams and **programming codes** as part of engineering projects, **scientific research** or educational activities and/or business activities.

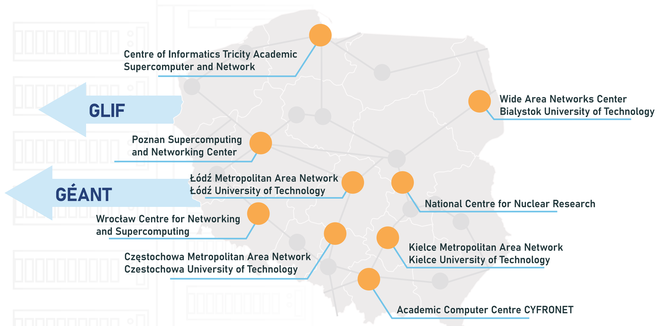
The NDS system is equipped with **interfaces** and mechanisms ensuring interoperability with **external environments,** including national infrastructures dedicated to data processing in HPC, Cloud (including PRACE-LAB), BigData and **AI models** (including PRACE-LAB2) and with other IT systems, as well as commercial clouds, incl. **AWS, Google or the National Cloud Infrastructure.**

DATA LAKE
- Dynamic data repositories
- The unstructured data processing
- Advanced analysis of historical and real-time

BIG DATA
- Effective use of large data volumes in business
- Integration of advanced Big Data, AI and ML solutions
- Support for small and medium-sized enterprises

EDGE COMPUTING
- Big Data processing from distributed sources
- High-level integration of the processing and storage services
- Real-time data processing

Distributed national infastructure



Infrastructure dispersion in 9 locations will be connected with a **fast fibre-optic network** in Poland and an effective connection with **GEANT and GLIF networks**.

Efficiency
TB/s

1 Exabyte
overall capacity

250 PB
tape libraries

250 PB
object storage

INFRASTRUCTURE

The NDS project provides a **hierarchical,** geographically dispersed, **redundant**, scalable, efficient and secure infrastructure for **data storage**, including servers, **server clusters**, specialized data storage systems, communication networks (local and wide-area), and **software**.

The infrastructure will be built based on **various architectures** of data storage systems, including classic server-matrix systems, tape libraries and modern, open **disk server clusters**, as well as various data storage technologies, including magnetic memories (tapes, disks), flash memory (**SSD, NVMe**) and non-volatile RAM (**NVRAM**), as well as data processing components (classic processors, dedicated accelerators, FPGA devices).

All systems will be combined in a hierarchical structure ensuring efficient, safe and **cost-effective data storage**, enabling effective data processing and management. The NDS infrastructure will use both classic and modern (**unconventional) system software stacks**, including cluster file systems and object storage systems.

SERVICES

The NDS infrastructure has embedded services that are typically implemented based on **mass data storage systems**, including backup, data recovery and mechanisms supporting **long-term** data storage and archiving, incl. management of data packing into archives, **archive metadata handling**, indexing and searching.

In addition, the NDS offers services and mechanisms supporting **applications** important from the point of view of academic institutions, research projects and commercial activities, the specific requirements of which can be addressed particularly effectively in a **large-scale distributed data** storage infrastructure.

These services include support for **multi-point storage** and access to data, **network buffering**, and mechanisms supporting the serving and distribution of **multimedia content**, incl. high-quality audio-video.

An integral part of the system are sets of functionalities typical for Data Lake systems related to searching, selective access to data, their **conversion between formats** as part of data presentation services or the automation of data processing based on defined events.

The NDS system infrastructure will integrate mechanisms supporting the construction of **data-centric Web applications**, allowing easy expansion and integration of new application functionalities with the data storage system through the use of modular and standardized components.

The access services developed as part of the project will also include **end-to-end data security** mechanisms, including cryptographic protection of the content and integrity of the stored data and non-repudiation of data operations.

DATA STORAGE
- Object-oriented
- File
- Hierarchical
- Reliable

DATA REPOSITORIES
- EOSC - Open Access
- Compatible with FAIR
- The unstructured data processing

BIG DATA ANALYTICS
- Artificial intelligence
- Data Lake model
- Cloud and HPC environments
- Edge computing

DATA SECURITY
- Data encryption
- Geo-replication
- Reliable storage with redundancy

FAIR DATA
Ensuring findability, accessibility, interoperability, and reusability

DATA REPOSITORIES
Collecting and gathering raw data into a structured data repository

METADATA
Attaching repositories to metadata aggregation and search engines

APLICATION DOMAINS
Reliable and sustainable data repositories adopted to domain needs

DATA STANDARD
Enforcing data standards and metadata descriptions

EOSC

Current challenges for science, combined with the policy of the European Commission, aim at more effective use of budgets supporting **open access** to research results and input data. This activity was called EOSC (European Open Science Cloud). An example of recent months is research on the development of a vaccine against COVID-19, conducted by **various data teams**. These types of activities are supported by the NDS.

We enable users to access our own **services and deposit data** in the NDS repositories. The fulfilment of these conditions will enable **data visibility** in an open forum in Europe and in the world.

The EOSC paradigms implemented in the NDS can be defined by **FAIR access** - Findability, Accessibility, Interoperability and Reusability.

BIAMAN · CZESTMAN · citask · CYFRONET · KIELMAN · Miejska Sieć Komputerowa LODMAN w Łodzi · NARODOWE CENTRUM BADAŃ JĄDROWYCH ŚWIERK · PSNC · WCSS

European Funds Smart Growth · Republic of Poland · European Union European Regional Development Fund