sciecominfo
Nordic-Baltic Forum for Scientific Communication

IMPLEMENTING NATIONAL OPEN ACCESS RESEARCH DATA ARCHIVE

Žibutė Petrauskienė, Saulius Maskeliūnas

Research data archives in Lithuania

There are number of different local registries, databases, and other repositories to store various research data in Lithuania. However, collection and storage of scientific data is often bound to internal regulations and procedures of certain institutions; access to scientific data in most cases is only available directly from the collecting institution. In addition, diverse software tools and various formats are used to store data; there is a lack of information about what institutions have collected what research data. Therefore, usage of this data in inter–institutional and multidisciplinary levels becomes rather complicated; there is a huge risk of losing highly valuable data. For those reasons, Vilnius University has decided to implement the project *National Open Access Research Data Archive* (in Lithuanian: *Nacionalinis atviros prieigos Mokslo Informacijos Duomenų Archyvas*, abbreviation: MIDAS).
The needs assessment was carried out in 2010, before starting the project. Scientific institutions were asked to fill a survey about existing practices in research data collection: what types of data they collect, what kinds of software they use, are they satisfied with current situation? During the initial stage of system design process in 2012, a similar survey was conducted once again, but now to determine the needs of potential users of MIDAS.

Implementation aims and principles of MIDAS archive.

The purpose of implementing National Open Access Research Data Archive is to establish the infrastructure that enables collection, organizing and storage of empirical and research data (with corresponding metadata), ensuring free, convenient, interactive search, access and analysis of data. MIDAS will ensure interaction and integration with other science, studies and biomedical data archives and data registers (i.e., Lithuanian Academic E-Library eLABa[1], Lithuanian Data Archive for Social Sciences and Humanities LiDA[2], Lithuanian Networked Digital Library of Theses and Dissertations Lit-ETD[3], National Medical Picture Archiving and Information Exchange System MedVAIS[4], etc.), compatibility with popular search engines (Google, Bing, Yahoo), and will especially support interdisciplinary research and collaboration both at personal and institutional levels. The users of National Open Access Research Data Archive will be researchers, lecturers, professors, students, science and studies institutions and/or their representatives, institutions wich present research data (e.g., hospitals) or their representatives, research and development (R&D) enterprises or their representatives, public administration institutions which use R&D statistical data, other interested physical and judicial persons. MIDAS archive will be based on usage of open code software, XML format and other open metadata, bibliographic, information retrieval standards (CERIF, CERIF for Datasets, CIF, DICOM, Dublin Core, MARC21, ISO/IEC 11179-1:2004, OAI-PMH, etc.). That will ensure compatibility with other information systems, data archives and registries in Lithuania and internationally (e.g., it is foreseen that MIDAS archive should comply with Data Citation Index[5] requirements of Thomson Reuters). The main development principles are: privacy and security (i.e. information confidentiality, integrity and non-repudiation), usability, accessibility (i.e., functioning 24 hours per day, 7 days per week), and extensibility (i.e., software architecture scaling in cases of incorporation of additional hardware).

[1] http://www.elaba.lt
[2] http://www.lidata.eu/en/
[3] https://etd.elaba.lt/
[4] http://accelerator.rssing.com/browser.php?indx=1133055&item=3051
[5] http://thomsonreuters.com/data-citation-index/

Lead institution of the National Open Access Research Data Archive project is Vilnius University[6]; project partner – Vilnius University Hospital Santariškių Klinikos[7]. The project participants are 13 science, studies and medical institutions (which have signed the collaboration agreements). The MIDAS project is funded by EU Structural Funds and national budget. Project budget is almost 15 mio LTL (~4.34 mio EUR); project implementation duration is 30 months; according to current plans, it should be completed in the 2nd quarter of 2015.

Planned MIDAS outcomes and peculiarities

At the MIDAS archive project the following components will be developed:

- Centralised infrastructure with the main and reserve data centres (located at the Centre of Information Technology Development[8] and at the Institute of Mathematics and Informatics[9] of Vilnius University), distributed virtualisation platform, Storage Area Network-type disk data storage, and hierarchical data storage for archive data (consisting of slower disk arrays and libraries of magnetic tapes);
- National united research data archive (with analytical software tools), which collects and stores empirical and other research data of different science areas;
- Infrastructure of biomedical research data collection and transferring (which will secure reliable accessibility and reuse of research data). Biomedical data component will consist of DICOM (for collecting data from medical equipment), ECG (for collecting electrical cardiogram data from medical devices), content management (for managing of collected data), data depersonalisation, and data archiving (for saving biomedical data in local data storage or central MIDAS database) modules;
- Public interactive e-service "Search, Delivery and Analysis of Research Data".

National Open Access Research Data Archive will collect data on researchers, R&D institutions, projects and grants, financing sources, classifiers of science and studies, MIDAS users, research objects, projects, equipment, data, and metadata.

MIDAS implementation will warrant the following possibilities:

- *Guaranteed safety and effective sharing of research data* among Lithuanian science and studies institutions and with international partners.
- *Increased quality of research outputs*: researchers will be able to explore and examine more data; easier access to research data will involve and encourage more academics and other people to participate in discussions about the accuracy, reliability and relevancy, etc. of scientific results.
- *Increased efficiency of research performance*: researchers will be prevented from duplication of effort in research data collection; time and material resources will be saved while collecting the same type of data from different institutions. In addition, time and material resources will be saved significantly by collecting data preserved in medical institutions and providing virtual access with no special software or hardware requirements, using ordinary web browsers (Internet Explorer, Google Chrome, Mozilla Firefox, Safari, Opera).
- *Increased variety of research outputs*: easier access to data will foster research that is unrealizable or economically unviable in practice due to consumptions of time and other resources for data collection; scholars, PhD students and others will have an interest in creating original and exclusive papers by formulating more diverse aims and objectives.
- *Decreased expenses for science and studies institutions*: lower expenses of research data storages, simpler scientometric analysis and research planning.
- *Increased quality of education*: lecturers, associate professors and professors will have more possibilities of exploiting the newest and most relevant data as well as be more informed while sharing knowledge and giving assignments for students.

Information infrastructure tools will be developed and implemented in MIDAS to ensure that scholarly

[6] http://www.vu.lt/en/
[7] http://santa.lt/
[8] http://www.ittc.vu.lt/en/
[9] http://www.mii.vu.lt/?lang=en

communication flows smoothly. Scholarly communication is considered as: (1) communication among researchers, lecturers, professors, (2) information sharing between science and educational institutions, and (3) dissemination of scholarly outputs both nationally and globally. Special attention in MIDAS will be given to motivating, encouraging researchers to present accurate and comprehensive metadata for their research data, and opening their data for usage by other researchers (e.g., increasing the quota of researcher's personal data space proportionally to his/her research data made public). Services and tools implemented in MIDAS will allow users to measure and analyse research data that are uploaded to the archive. In addition, data analysis subsystem will allow to process research data with various multi-dimensional data analysis, visualisation, classification and grouping algorithms, allowing users to run large distributed and parallel computations using Vilnius University grid and supercomputer resources. The data analysis process involves: (1) user authentification based of MIDAS single sign-on (SSO), (2) selection of data which is stored in MIDAS for analysis, (3) defining the workflow of data analysis, selecting the steps from the following set: data getting, initial data preparation (cleaning, filtering, transformation, transposition, separation), selection of data analysis algorithm and setting of control parameters, appointing calculation resources, review of analysis results, review of technical characteristics of data analysis algorithm operation, saving of analysis results in MIDAS infrastructure or in user's computer. The interface of data analysis tool will be realised as webpage, user will use it with web browser; any additional installations in user's computer and any programming skills will not be needed.

Conclusion: MIDAS will provide virtual services for researchers and other participants in research and education that can lead to more efficient, effective and higher quality research. Users will have the possibilities to register, find and cite research data, search for other infrastructures or tools (which provide data archiving services), use them, also share or integrate data and tools to other science and studies infrastructures. In addition, National Open Access Research Data Archive will increase the visibility of Lithuanian science in international context and international cooperation possibilities, because of simpler, more convenient, unified, advanced possibilities of research data collection, analysis, application and sharing.

Dr. Žibutė Petrauskienė, Head of Scientific Information Data Department, Vilnius University Library, MIDAS Project Coordinator, zibute.petrauskiene@mb.vu.lt

Dr. Saulius Maskeliūnas, researcher and vice-director at the Vilnius University Institute of Mathematics and Informatics, senior specialist of MIDAS project, saulius.maskeliunas@mii.vu.lt