BUILDING AN INFORMATION MANAGEMENT AND EXPLOITATION SYSTEM THAT INTEGRATES ONLINE SCIENCE AND TECHNOLOGY DATABASES

Vu Thi Khanh Trinh[2], Pham Ngoc Minh[1], Tran Huy Duong[1],*

Nguyen Long Giang[1], Hoang Thi Thu Dung[2]

[1]Institute of Information Technology - Vietnam Academy of Science and Technology

[2]Tan Trao University, Vietnam

[*]Email address: pnminh@ioit.ac.vn

DOI: 10.51453/2354-1431/2023/959

Article info	Abstract:
	The database of science and technology plays a crucial role for each country and region as it stores the scientific research achievements and technological developments of each country. In the current trend of digital transformation, countries have been investing strongly in digitization and database establishment for archiving and managing scientific data and technological development to serve their socio-economic development and national security and defense. Therefore, the authors propose to build an online administration and exploitation information system that integrates information mined from scientific and technological databases with the goals of standardizing scientific and technological data and building a dedicated interconnected system of relevant scientific and technological databases.

XÂY DỰNG HỆ THỐNG QUẢN TRỊ, KHAI THÁC THÔNG TIN TÍCH HỢP CƠ SỞ DỮ LIỆU KHOA HỌC & CÔNG NGHỆ TRỰC TUYẾN

Vũ Thị Khánh Trình[2], Phạm Ngọc Minh[1]*, Trần Huy Dương[1],

Nguyễn Long Giang[1], Hoàng Thị Thu Dung[2]

[1]*Viện Công nghệ thông tin – Viện Hàn lâm Khoa học và Công nghệ Việt Nam*

[2]*Trường Đại học Tân Trào, Việt Nam*

**Địa chỉ email: pnminh@ioit.ac.vn*

DOI: 10.51453/2354-1431/2023/959

Thông tin bài viết	Tóm tắt
	CSDL về KH&CN đóng vai trò quan trọng đối với mỗi quốc gia, vùng lãnh thổ vì nó lưu trữ các thành quả nghiên cứu khoa học và phát triển công nghệ của mỗi quốc gia. Trong xu thế chuyển đổi số hiện nay, các quốc gia đã và đang đầu tư mạnh mẽ cho việc số hóa và xây dựng các CSDL lưu trữ, quản lý dữ liệu khoa học và phát triển công nghệ phục vụ phát triển kinh tế - xã hội, đảm bảo an ninh, quốc phòng. Vì vậy nhóm tác giả đề xuất xây dựng một hệ thống quản trị, khai thác thông tin tích hợp CSDL KH&CN trực tuyến với mục tiêu chuẩn hóa dữ liệu KH&CN và xây dựng hệ thống liên quan kết nối giữa các CSDL KH&CN.

1. Introduction

In recent decades, along with energy and resources, science and technology (S&T) have been regarded as the third potential factor that plays a vital role in fostering the development of all human social activities. Scientific research and technological development have long been highlighted and invested in around the world, particularly in developed countries, and have become one of the most important resources in the economic-social development, ensuring national security. The fourth industrial revolution is currently taking place, with a significant and profound impact on various aspects of socioeconomic life, ushering the globe from the industrial age to the information and knowledge-based economic development age. In this trend, the field of S&T is undergoing a significant revolution, with S&T information systems and databases being constructed and operated internationally, in each country and region.

2. The role of S&T databases in socio-economic development in the world

S&T databases play an important role for each country and region since it stores the outcomes of their scientific research and technological development. In the present digital transformation trend, countries have been investing extensively in digitizing and building databases to store, manage scientific data, and developing technologies for socioeconomic development, as well as ensuring security and national defense. Therefore, the development of S&T database management and exploitation systems needs to focus on the standardizing S&T data and building a related system connecting the S&T databases.

S&T databases generally include: 1) electronic publication databases, including free databases such as open-access journals or premium databases such as Proquest, ScienceDirect, Springerlink, IEEE..; 2) indexing and citation information databases such as ISI, Scopus, PubMed, Google Scholar, OpenCitation..; 3) open-access repositories such as Datacite, arXiv, OpenAIRE..; 4) registration and identification management systems for published documents and researchers, such as Crossref (DOI), ORCID, ISNI..; 5) specialized research information systems of organizations such as PubChem (chemistry), GlobalChange (climate change), InspireHEP (high-energy physics); 6) databases on S&T tasks; databases on S&T organizations and experts; and 8) databases on inventions and patents. The following are some examples of typical applied studies on the condition of building and developing S&T databases in general and S&T databases in some countries throughout the world.

2.1. Japan

Japan established the Japan Science and Technology Agency (JST) [1], which is responsible for collecting, storing, processing, and disseminating S&T information in Japan. JST has created a S&T information system comprised of four major databases, which are as follows:

-Database of scientific research organizations: name of the organization, address, history of development, research fields, main activities…

- Database of research experts: name of the expert, organization, title, research fields, research results..

- Database of topics and tasks of scientific research and technology development: name of the topic, lead researcher, topic description, research fields, research results..

- Database of research resources: name, contact address, resource description, usage environment..

2.2. Malaysia

The Malaysian Science and Technology Information Center (MASTIC) [2] has developed the MASTICLink portal, which stores and provides information resources on S&T human resources, projects and organizations. MASTICLink includes the following component databases:

-S&T Human Resource: a database of individuals involved in S&T activities in Malaysia. The database is divided into 3 parts: S&T human resources; S&T experts in each field; members of S&T academic associations;

-S&T Projects: a database of ongoing and completed research and development projects. The projects are conducted by universities and research institutes and are sponsored by the Malaysian Ministry of Science and Technology through S&T development funds, technology innovation funds, etc.;

- S&T Organizations: a database of S&T organizations in Malaysia, classified by industry, field, and type of organization (governmental, private, non-profit);

- S&T Facilities and Equipment: a database of laboratory machinery and equipment, and other facilities available for research and development in Malaysia;

- TechMart: a database of research and innovation results that can be commercialized.

2.3. China

The Institute of Scientific and Technical Information of China (ISTIC) [4] has also developed the Chinese Scientific Institution Database (CSI) including information on over 10,000 research and development organizations in various fields; and the Who's Who database which includes information on over 4,000 officials involved in S&T policy-making and management in China.

Through our research and survey, we have identified two issues that need to be addressed: (1) Inefficient operation due to lack of funding, lack of data linkage between agencies, lack of control, and inefficient exploitation of domestic and foreign S&T activities; (2) Lack of skilled personnel in charge of information exploitation, analysis and synthesis.

3. Overview of the National Database on Science and Technology

The Ministry of Science and Technology (MoST) has developed and issued mechanisms and policies to promote digital transformation in the S&T field. On June 28, 2017, the MoST issued Circular 10/2017/TT-BKHCN [10], which stipulates the construction, management, operation, use, maintenance, and

development of the national S&T database. Based on Circular 10/2017/TT-BKHCN, the National Science and Technology Information Center under the MoST has built and operated the national database on science and technology to unify the management of all data on S&T resources in real time, meet the requirements of integration and sharing, and promote the application of S&T methods in data collection, management, use, and exploitation in the S&T field. The National Database on Science and Technology ensures compliance with the architecture of the National Information System, meets the standards for databases, and complies with the technical standards and regulations for information and communication technology. It also meets the demand for accessing and exploiting information and data to serve the construction of strategies, policy planning, management, and operation in the field of state management of S&T, ensuring timeliness,

completeness, transparency, and compliance with international practices.

The National Database on Science and Technology consists of 10 component databases:

- Database on S&T organizations;

- Database on S&T research and development personnel;

- Database on S&T tasks;

- Database on scientific publications and citation index;

- Database on S&T statistics;

- Database on technology, high technology, and technology transfer;

- Database on S&T information in the region and the world;

- Database on S&T enterprises;

- Database on intellectual property information;

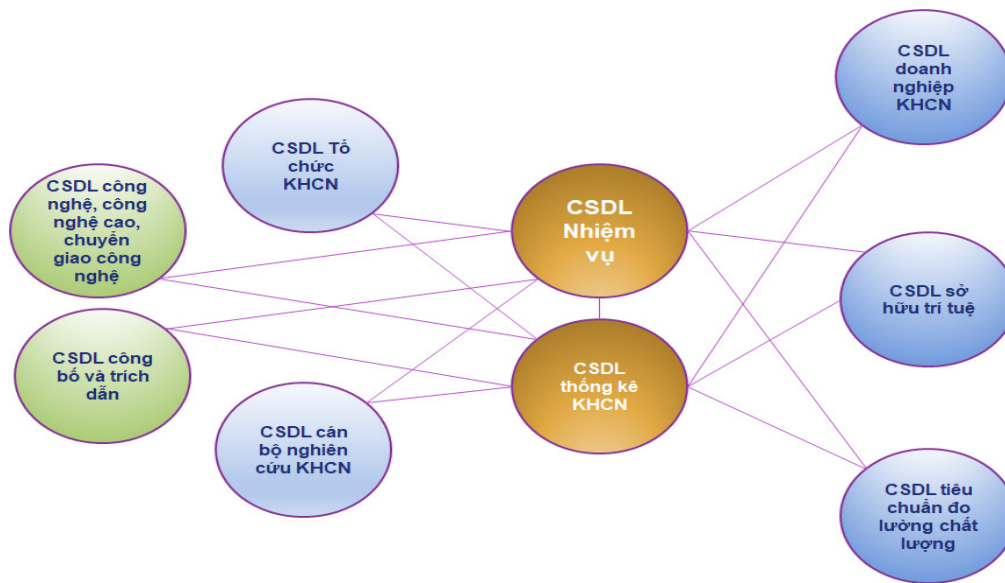- Database on quality measurement standards.

Figure 1: National database on science and technology

In addition to the aforementioned databases, there are also other types of scientific and technological data, such as assets formed during research, laboratories, innovative and technological transfer activities, barcode numbers, nuclear radiation safety, and open data on science and technology. These types of data also need to be managed and stored uniformly in a centralized database.

Given the requirements for connectivity and information exchange with the national database mentioned above, building S&T databases and software for updating, storing, and exploiting S&T information, especially the three categories of

information (information about S&T organizations, information about research personnel, and information about S&T tasks) is a crucial requirement for provinces, departments, and cities in the current situation. Until now, some provinces and centrally-governed cities have built and operated S&T databases and software systems to connect and exchange data with the national database of S&T. The current form of data connection and exchange mainly involves updating information on the three categories mentioned above on the national database of S&T using provided accounts, but automatic connection and exchange of information

with the national database of S&T through APIs has not yet been implemented.

4. Building a model for online management and exploitation of the database on science and technology

4.1 System architecture

In the context of national digital transformation, various fields and industries are strengthening the construction of digital projects and specialized databases, creating specialized data infrastructure and integrating, linking with the data infrastructure of the province. The field of science and technology is not outside this trend. Therefore, digitizing and building a database and a software system for managing the database on S&T is essential to create a digital data bank on S&T, integrate and exchange information with the national database on S&T and other databases in the province. The digital database and management software are built on a modern technology platform with large data to meet the requirements in the new situation of online implementation, suitable for development trends. In response to this demand, the Institute of Information Technology - Vietnam Academy of Science and Technology is building a system for online management and exploitation of the database on S&T (currently being implemented in some provinces and cities nationwide such as Ninh Thuan, Vinh Phuc..) with the following general requirements:
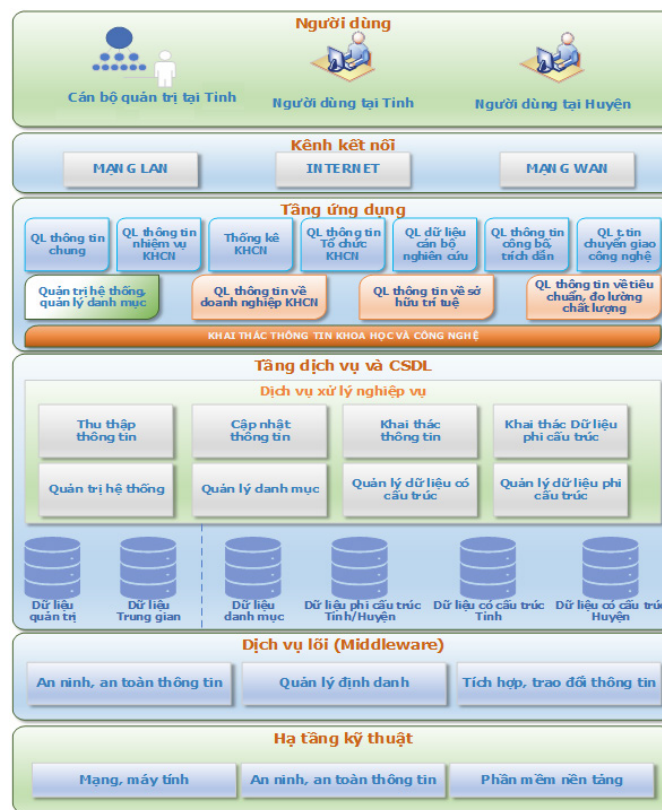

Figure 2: Architecture of the management and exploitation system for online S&T databases

The software meets the new business requirements for managing the national database of online S&T according to the draft circular on managing the national database on S&T of provinces, cities, and the MoST.

The national database on S&T includes the following main blocks:

1) Data warehouse of S&T: includes component databases and conversion tools from databases, component systems, and systems managed by other units to the data warehouse.

2) National S&T database management and exploitation modules:

- Data update and management subsystem: allowing user authorization, updating of new indicators without data sources, synthesis, and approval of data into the warehouse. The system completely renews the data collection and updating work to maximize data acquisition, aiming at users. The units generating data

will participate in providing, updating, and approving input data.

- Data exploitation subsystem: Implemented according to the intelligent report model supporting decision-making (Business Intelligence - BI)

- S&T publication and information dissemination subsystem: Published on the website/portal for citizens and businesses to exploit.

3) Real-time data linkage, integration, and sharing module:

- A connection and data platform sharing with the Vietnam Data Exchange Platform, and other national information systems and databases.

- Providing APIs to allow information systems and component databases of units under the MoST to provide data to the data warehouse.

- Provide APIs to allow information systems and softwares of the MoST ant the subordinate units to connect, extract, and use data in the national S&T data warehouse.

-The databases and softwares are built on a modern and advanced technology platform, ensuring security, information safety, scalability, and integration with other systems through connection standards, using connection technology via APIs.

-The S&T database of the unit can connect and exchange data with the national S&T database to meet the requirements in Circular 05/2021/TT-BKHCN on technical requirements for input data of the national S&T database, including three basic information groups: information on S&T organizations, information on S&T research personnel, and information on S&T tasks.

-The S&T database management and exploitation system meets the requirements for different user groups to quickly, timely, and accurately exploit information.

4.2. Data architecture model

The data architecture model of the S&T database system applies a new model based on large data architecture as follows:
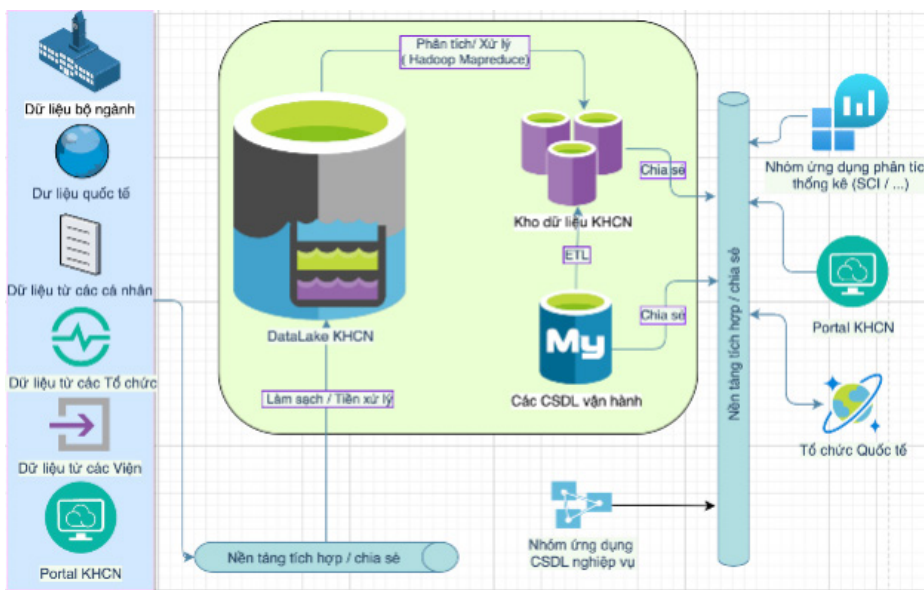

Figure 3: Data architecture model of S&T database system

The functional blocks of the model are as follows:

1) Integrated/Shared Platform Block for System Input:

- Input: Data collected from various sources including shared data from industry departments (including both original and corrected data), data from S&T departments of 63 provinces and cities, shared data from other organizations (countries that share

S&T information, companies and individuals, research institutes applying S&T). Data can be in the form of Text File, CSV, PDF, and other text formats or in the form of traditional databases.

-Output: Data through this block will be put into the Data Lake.

2) S&T Data Lake: This is the storage and analysis/ processing location for all the collected data in the

large database (including both normalized data and unnormalized data such as text files, CSV, audio, images, streaming files, etc.).

-Input: Data collected and processed in the previous step is sent to the Data Lake.

- Output: Analyzed and processed data is sent to the S&T Data Warehouse.

3) S&T Data Warehouse: This is also a Big Data database, but it has been analyzed and organized clearly from the data of the Data Lake in the previous blocks. The databases in this block are an upgrade from the data of the National S&T Database in the previous stages.

- Input: Analyzed and processed data from the operational database of S&T applications.

- Output; The output data from this block can be used through the following three blocks: directly shared through the integrated/sharing platform with specialized analysis/prediction/statistics applications for specific fields; shared with external partners through the integrated/sharing platform, including research institutes, individuals, businesses, and foreign partners who desire to share and are authorized to share information about S&T data; and shared with specialized business applications groups for data exploitation that is not part of the operational database.

4) Specialized operating database block for S&T: This block will serve the business problems of the National Agency for S&T Information in the Operational Business Applications Group.

-Input: Data from discrete databases in the previous stages are migrated, and data generated during the operational business process.

-Output: The data in this block will be used for the following two blocks: Main database of some specialized business applications and part of the input of the S&T Data Warehouse Block. It is shared with some other parties through an integrated/sharing platform (National S&T Information portal).

5) Application software/tools block: This block places software/tools developed by the Agency and commercial software to exploit the S&T Data Warehouse and operating databases.

6) Sharing/integrating output block of the System: This block is placed in the DMZ of the data center. The main task of this block is to share data for external applications such as units of other ministries, external organizations.

-Input: Data from specialized data warehouses and business databases.

-Output: The National S&T Information Portal, and external databases that intend to use S&T data.

4.3. Deployment Model

Citizens, organizations, and businesses can access the system through the Internet to search for information, retrieve data, and exploit data. Related operational units access the system through the Internet or a dedicated line to the service provider's hosting system.
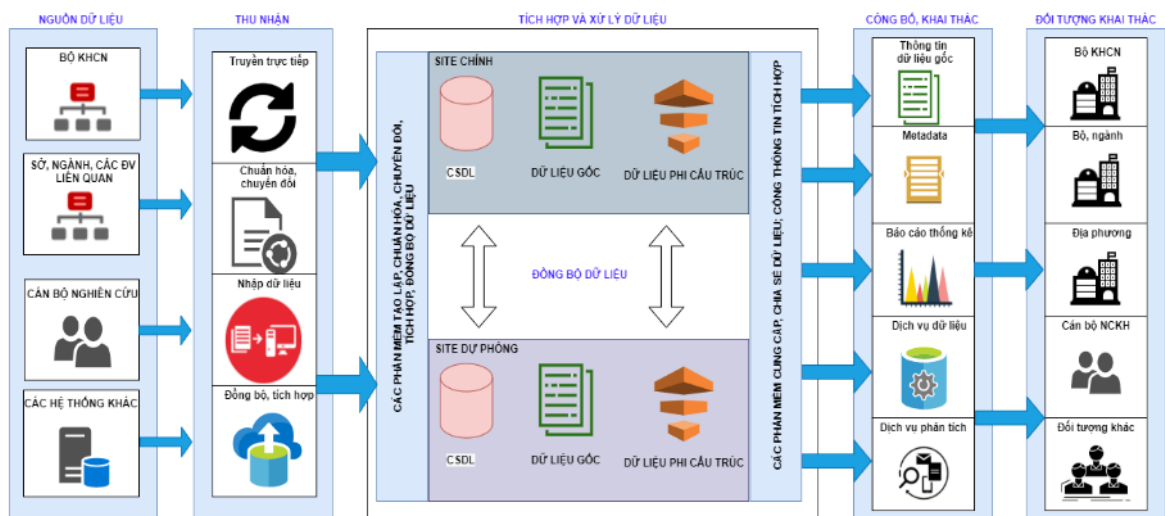

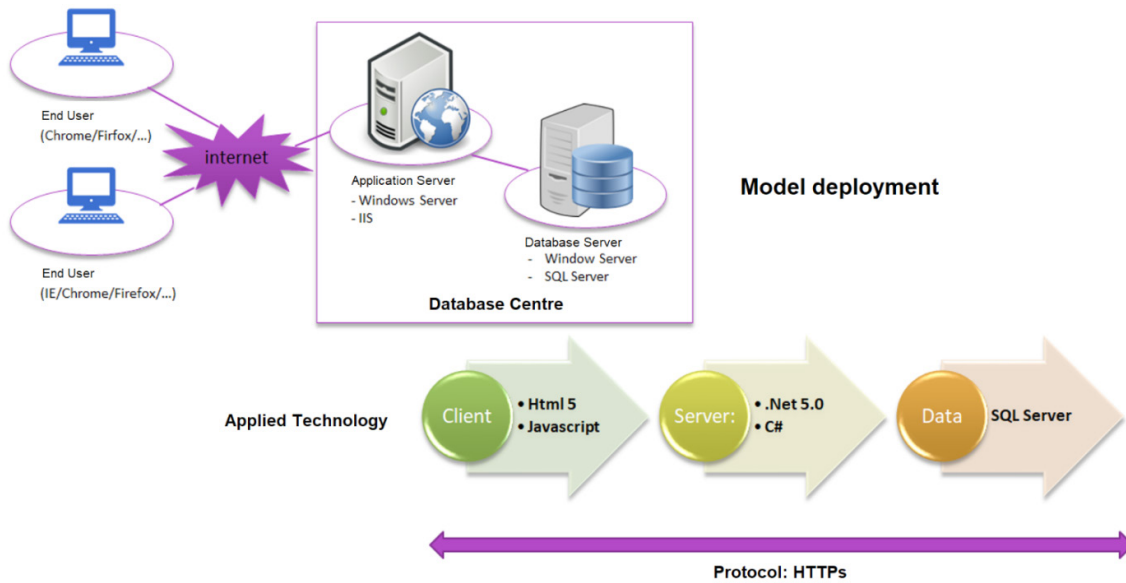**Figure 4: Deployment process model**

Figure 5: National S&T Information System Infrastructure

The deployment model of the system consists of the following main components:

Database management system: The database management system for the national S&T information database is MySQL (or PostgreSQL) and MongoDB. This management system meets the criteria of portability, modernity, and use of the latest technologies.

Operating system: The operating system for the national S&T information database is Win Server for the database servers (data lake, data warehouse) and data integration.

Integrated and data sharing software: The national S&T information database selects the C-dat platform for its advantages, including a large development community, extensive support, modern technology, and cost-effectiveness in today's strong application of information and communication technology.

Data warehouse platform: The database is built according to the standards of data warehouse systems, with a star schema (data warehouse star) model that stores all data sources related to the allocation of S&T capital, capital utilization, projects, topics, project applications, and the effectiveness of S&T projects. Specialized data marts are designed to meet the information retrieval needs of users.

Application development platform: The national S&T information database uses Java/.NET platforms as application servers for application software.

Conclusion

The National Database on Science and Technology is considered an important data infrastructure in the field of Science and Technology, providing a platform for connecting and integrating with component databases. It not only contributes to providing and sharing data information with the systems of the Ministry and Government but also helps organizations, businesses, and citizens interested in accessing data and information on Science and Technology, contributing to increasing transparency of information and supporting research, development, and production in the field of Science and Technology in Vietnam.

At the same time, the S&T Database system creates a research map as a multifunctional research information platform including Vietnamese and international researchers. The platform provides tools for registered researchers to easily collect information, including research papers and other works. Registered researchers can create their own personal page on the research map to store their CV and research works. Moreover, the research map serves as a virtual laboratory and an online research community for sharing and generating ideas for future collaborations.

References

[1]   https://www.nii.ac.jp/publications/kaken/HTML1999/99Kikuc-E.html.

[2]   https://mastic.mosti.gov.my/information/introduction

[3] https://stii.dost.gov.ph/.

[4] https://www.istic.ac.cn/isticcms/html/1/istic-en/en_index.html.

[5] https://aassrec.org/member-thailand-national-research-council-of-thailand-nrct/.

[6] Decision No. 749/QD-TTg dated June 3, 2020 of the Prime Minister approving the "*National Digital Transformation Program until 2025, with a vision towards 2030*".

[7] Decision No. 27/QD-UBQGCDS dated March 15, 2022 of the National Committee on Digital Transformation on the issuance of the *"Action Plan of the National Committee on Digital Transformation for 2022*".

[8] "*Digital Transformation Handbook*" by the Ministry of Information and Communications, 2021.

[9] Circular No. 05/2021/TT-BKHCN dated June 17, 2021 on "*Technical requirements for input data of the national database on science and technology*".

[10] Circular No. 10/2017/TT-BKHCN dated June 28, 2017 on "*Regulations on the construction, management, exploitation, use, maintenance and development of the national database on science and technology*".

Acknowledgements