https://doi.org/10.26069/greynet-2025-000.765-gn

# Global information Repository Research for STI Development: KISTI-GreyNet Collaborative Research Project Report

*Dominic Farace*
*GreyNet International*

*Jayhoon Kim and SeokJong Lim*
*Korea Institute of Science and Technology Information*

## Introduction

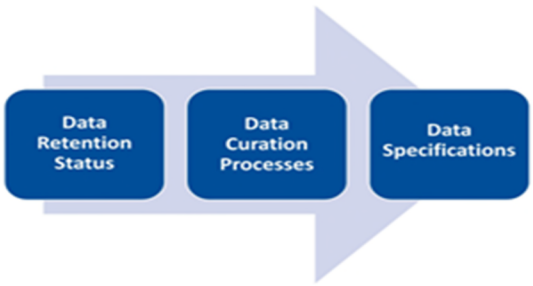
The Korea Institute of Science and Technology Information (KISTI) and GreyNet International collaborated in a joint research project entitled 'Global Information Repository Research for STI Development'. GreyNet an established community of practice in the field of grey literature for over thirty years has contacts with research organizations in more than 40 countries worldwide. GreyNet brings to this project its expertise in the promotion and management of scientific and technical information outside the sphere of commercial publishing. This six-month collaborative research project was launched in the first week of April 2024 and closed the final week of September 2024.

## Project Objectives

The goal of the research project is to secure curated and diverse types of data held by organizations worldwide and in so doing expand the data spectrum through collaboration between AI and humans. GreyNet was chosen by the Korea Institute of Science and Technology Information (KISTI) for this joint research project because of its international scope, sectors of information and organizational contacts, diverse range of document types in digital and non-digital formats, its sustained information resources, and open access compliance. On 29 April 2024, the first of many ZOOM meetings was held in which both KISTI and GreyNet presented an understanding of their roles in this joint research project[1].

## Method of Approach

This six-month project was carried out in three phases each of which incorporated a questionnaire and a controlled survey population.




## Phase 1: Data Retention Status

An online questionnaire on the Data Retention Status[2] of an organization was held. The selection of the target survey population was carried out by GreyNet using the 4th edition of the International Directory of Organizations in Grey Literature (IDGL)[3], which contains 221 organizational records having a ROR-ID. The number was then reduced by the currentness of the

https://doi.org/10.26069/greynet-2025-000.765-gn



record being between 2022 to 2024, which then accounted for 91 organizational records. A selection from this set of records was then made based on the organization's affiliation with GreyNet such as associate membership, number of contacts within any one organization, number of publications published by or in collaboration with GreyNet, editorial affiliation with GreyNet, an organization's Funder-ID, and previous involvement in a research project together with GreyNet. The number was finally reduced to 51 records excluding the KISTI and GreyNet records, which then accounted for the target survey population in the initial phase of the study. The online questionnaire consisted of ten items – nine multiple choice and one open-ended.

| Survey Population | Number of Survey Recipients | Number of Survey Respondents | Survey Response Rate |
|---|---|---|---|
| Controlled | 51 | 25 (49%) | 89% |

In summary the questions dealt with • Issues related to the types of data produced, collected, and if paid for by one's organization. • The metadata assigned and managed in a bibliographic and/or full-text database along with internet online services provided. • The type(s) of grey literature one's organization collects, manages, and provides for a fee, free-of-charge, or by CC licensing. • The final question requested the respondent's name and email address, thus confirming the controlled nature of the survey population.

**Survey Results**

Responses to the questionnaire were then recorded by way of a Coding Sheet and analyzed. The findings indicate that:

(1.) The types of data both produced and collected by the organizations are statistically comparable;

(2.) Academic journals account for a near half of the acquisition, subscription, and usage fees of an organization;

(3.) Regardless of the types of grey literature, the metadata assigned and managed in a database is near proportionate for all of the types;

(4.) A correlation appears in the types of grey literature that are housed full-text, online accessible, and CC licensed or otherwise open access compliant;

(5.) On average, 40% of the grey literature acquired by the organizations are free of charge. While on the other hand, reports of various types account for 40% of the organizations purchase fees;

(6.) Of the 25 organizations responding to the survey, only five percent do not provide grey literature free of charge.

The results of the survey remain compiled and preserved in SurveyMonkey as well as in the DANS, Data Station for the Social Sciences and Humanities[4].

https://doi.org/10.26069/greynet-2025-000.765-gn



**Phase 2: Data Curation Processes**

An online questionnaire on the Data Curation Processes[5] carried out by the organization was conducted. The 25 respondents to the questionnaire on data retention determined the population of the questionnaire on Data Curation. Three of those 25 respondents answered anonymously, leaving 22 survey recipients to form the survey population in this second phase of the study on Data Curation Processes. The online questionnaire consisted of ten items – nine multiple choice and one open-ended.

| Survey Population | Number of Survey Recipients | Number of Survey Respondents | Survey Response Rate |
|---|---|---|---|
| Controlled | 22 | 14 (63,6%) | 93% |

In summary, recipients were asked to • Indicate the area(s) of greatest opportunity in reach of your organization and the area(s), where your organization currently shows the greatest opportunity. • Indicate the area(s) closest to the function you carry out within your organization and the critical task(s) of data curation in which your organization is involved. • From a user perspective, indicate key challenges to improve the user-experience in your organization and what priorities regarding data curation would improve the user experience. • Indicate areas of interest in community building, collaboration on data sharing, and technical cooperation that would be of benefit to data curation. • What expected change(s) would take place, when AI technology is introduced in the existing work environment; and, in what area(s) would you make recommendations to improve the data curation process in AI environments. • The final question requested the respondent's name and email address, thus confirming the controlled nature of the survey population.

**Survey Results**

Responses to the questionnaire were then recorded by way of a Coding Sheet and analyzed. The findings indicate that:

(1) Research and information service functions are seen as the greatest opportunities of an organization.

(2) While the production and preservation of the collections are equally seen as an organization's strengths, the current greatest strength is the innovative user experience.

(3) While data collection and retention are equal in their roles carried out by the survey respondents, data management and data sharing are the closest to the roles they now carry out within their organization.

(4) Data conversion and data innovation (such as machine learning and AI services) are among the critical tasks in one's organization.

(5) Data discovery, data use, and reuse are together the key challenges in improving the user experience.

(6) Policy direction embedded in a strategic plan, where AI technology is incorporated in data processing, is a priority in the curation of data that would improve the user experience.

https://doi.org/10.26069/greynet-2025-000.765-gn



(7) Other areas that would highly benefit data curation are data sharing and the sharing of AI technology.

(8) When AI technology is introduced in the existing work environment, the greatest expected
change is the shift from face-to-face service to chatbot and technical system responses.

The results of the survey remain compiled and preserved in SurveyMonkey as well as in the DANS, Data Station for the Social Sciences and Humanities[6] .

**Phase 3: Data Specifications for Data Sharing**

An open-ended questionnaire dealing with Data Specifications for Data Sharing[7] by an organization was emailed**.** The population in this third and final phase of the project was determined by KISTI from among the 14 respondents to the questionnaire on Data Curation of whom 6 were selected and to which a 7th was included outside the respondents to the survey.

| Survey Population | Number of Survey Recipients | Number of Survey Respondents | Survey Response Rate |
|---|---|---|---|
| **Controlled** | **7** | **6 (86%)** | **100%** |

In summary, recipients were asked to enter contact information, key shared data types, and a description of each shared data type:

- Origin of the data type: (Brief description of the data origin and its creators)
- Source of information: (Department and contact information)
- Current holdings: (Amount of data)
- Data update and provision cycle: (Weekly, monthly, etc.)
- Service homepage: (URL)
- Data processing and activities: (Workflow, procedures in brief)
- Metadata fields and elements:
- Data recording period and scope: (Year of construction)
- Data Sharing Terms and License Policy

The final question inquired if there was further interest to collaborate in data sharing for purposes of research and public awareness?

**Survey Results**

Only five of the six survey respondents had an interest in further collaborating in data sharing given certain conditions. In brief, these five account for the coverage of 8 sustained information resources, consisting of 14 data types within 10 disciplines/fields with holdings of over a half million digital objects and 1.2 million metadata records, spanning the years 1980 to 2024 in addition to other older retrospective input with diverse CC and other forms of licensing.

The responses to the questionnaire were then entered in a Table (see below) showing across-the-board totals.

https://doi.org/10.26069/greynet-2025-000.765-gn

**Data Specifications for Data Sharing**

| Respondent ➔ | 1 | 2 | 3 | 4 | 5 | In Total: ↓ |
|---|---|---|---|---|---|---|
| **Resource(s):** | 2 | 3 | 3 | 1 | 1 | **8 Resources** *(Two of which overlap)* |
| **Data Types:** | Research Reports<br>Yearbooks<br>Statistics<br>Research Papers<br>S&T Terms | Videos<br>Research Data<br>Conference Papers | Research Reports<br>Research Papers<br>Images | Research Reports<br>Statistics<br>Research Papers<br>Reference Databases<br>S&T Terms<br>Systematic Reviews<br>Standards and Guidelines | Videos<br>Sound Recordings | **14 Data Types** *(Some of which overlap)* |
| **Disciplines/Fields:** | Aquatic Sciences<br>Fisheries | Library<br>Information Science<br>Publishing<br>Multidisciplinary | S&T<br>Research Management | Life Science<br>Health Care | S&T<br>Multidisciplinary | **10 Disciplines/Fields** |
| **Holdings:** | 180,000 | 667 | 260,260<br>---<br>1,2M metadata | 13,200 | 47,000 | **501,127 Objects**<br>---<br>**1.2M Metadata** |
| **Period Covered:** | 1980s – 2024 *(including retro-records dating back 100 years)* | 1994 – 2024 | 2009 - 2024 | 1999 - 2024 | 2014 - 2024 | **1980 – 2024** *(including retro-records dating back 100 years)* |
| **Licensing:** | CC - Licensed | CC - Licensed | | Per terms of the provider | CC - Licensed | **CC As well as other licensing** |
| **Interest in Collaborating:** | Positive / Conditional | Positive / Conditional | Positive / Conditional | Positive / Conditional | Positive / Conditional | **Positive / Conditional** |

## Follow-up and Closing Statement

The expected outcome of this joint research project was to move from a research perspective to an implementation perspective in the collection of new data via collaboration with select grey literature organizations. It is important to learn more about the specific uses and applications of the shared data as well as agreement as to which resources are eligible. In follow-up, it will be necessary to pursue contact with the organizations that participated in all three questionnaires to discuss issues related to sharable content, licensing, a memorandum of understanding, data delivery methods, and the data sharing process. It is in this way that the conditions linked to their interest in collaborating will be addressed. And, it is to this end that the implementation of diverse shared data types can be achieved among their combined networked resources in accordance with FAIR data principles.

Our joint research project has been able to identify a short-list of recognized organizations in the field of scientific and technical information. By way of data sharing, they can bring together diverse and openly accessible data types for purposes of research and public awareness. Incorporating their key shared information resources via a network platform will allow for viable responses to problems and issues challenging science and society. This network platform likewise holds real potential for further expansion and development.

https://doi.org/10.26069/greynet-2025-000.765-gn

## Referenced Links

[1] https://www.greynet.org/images/Appendix_2_GreyNet_-_KISTI_Presentation_April_29,_2024_.pdf

[2] https://www.greynet.org/images/Appendix_4_Questionnaire_-_Data_Retention_Status.pdf

[3] https://www.greynet.org/images/IDGL_2024.pdf

[4] https://ssh.datastations.nl/dataset.xhtml?persistentId=doi:10.17026/SS/EJ1YUF

[5] https://www.greynet.org/images/Appendix_10_Questionnaire_-_Data_Curation_Processes.pdf

[6] https://ssh.datastations.nl/dataset.xhtml?persistentId=doi:10.17026/SS/G0LHGY

[7] https://www.greynet.org/images/Appendix_15_Questionnaire_-_Data_Specifications_for_Data_Sharing.pdf