Twenty-Fifth International Conference on Grey Literature
'Confronting Climate Change with Trusted Grey Resources'
OBA Congres • November 13-14, 2023
Oosterdokskade, Amsterdam, Netherlands

https://doi.org/10.26069/greynet-2025-000.797-gn

A conceptual design case study of an on-demand scientific information collection based on AI system

Jayhoon Kim, Miwhan Hyun, and Hyejin Lee
KISTI. Digital Curation Center, Division of National S&T Data. 245 Daehak-ro, Yueseong-gu, Daejeon, Rep of Korea

Twenty-Fifth International Conference on Grey Literature
'Confronting Climate Change with Trusted Grey Resources'
OBA Congres • November 13-14, 2023
Oosterdokskade, Amsterdam, Netherlands

A conceptual design case study of an on-demand scientific information collection based on AI system

2023.11.14. Jayhoon Kim, Miwhan Hyun, and Hyejin Lee

KISTI Digital Curation Center

1. Background

❑ Need for an on-demand scientific information collection process design

Researcher's information needs

- Need to respond to the up-to-date sci & tech information demand of researchers
- Need to expand content line up in regards to 4th research paradigm
- User needs for on-demand content production and provision methods that respond to various levels of scientific and technological information needs.

3

2. Research Purpose

❑ User-centric digital curation process re-design fits for 4th research paradigm

Broadening information type | Up-to-date | Trendy | On-demand | Re-usable

AS-IS	TO-BE
Traditional science and technology content such as	Global science and technology trends , news Expansion of information sources, etc.
Providing customized content corresponding to	Providing content in a customized format
Customized content in the form of reports is produced in small quantities by humans.	A large amount of customized content can be produced using the human-in-the-loop method.

4

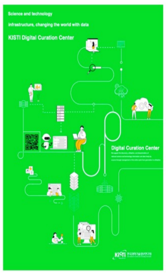
3. Research methods

❑ A theoretical model design based on current status analysis and literature review

AS-IS, TO-BE Analysis
- KISTI's current digital curation policy, process review
- Directions for new process

Literature research
- Researcher information needs
- Content processing related AI technology
- Copyright related regulations/issues

Process design
- Data collection
- Data processing
- Data quality management



5

4. Results (1)

❑ User-centric digital curation process re-design fits for 4th research paradigm

AS-IS, TO-BE Analysis

	AS-IS	TO-BE
Data source	Based on research results (papers , reports , patents, etc.)	Expanding variety of sci&tech information (Scientific publisher, website, social media)
Data recency	Low (Longer period of review process papers , patents, etc.)	High (Trendy sci&tech information, preprints , science news, etc.)
Data usability	Search, link to fulltext files to get research idea	Reusable centric, machine application (Reproduce, translate, summarize, analyze)
Service level	Data supplier-centric (search, browse)	Data user-centric (On-demand, usable format, generative)
Secure copyright	Passive (License information management)	Active (Open Access movement, License agreement)

6

4. Results (2)

❑ Data source research

Major publishers
- Traditional publishers
- OA publisher

Science websites
- KISTI's current digital curation policy, process review
- Directions for new process

Open access portals
- UK JISC CORE
- OpenAIRE
- Unpaywall, etc.

Social Media
- ResearchGate, Korean scholar communities(Specialized research information centers)
- X (Twitter), Facebook, etc.

Current articles
- Pre-print
- Articles in Press

7

4. Results (3)

❑ Data sourcing strategy

Data type	Strategy
Open data	• Open access papers can be used free of charge , but it is necessary to check the license conditions for each paper in order to produce secondary works such as paper translation and summary. • The license notation for open access papers often uses the Creative Commons License (CCL) . • In relation to this project, the use of derivative works must be possible, so a license excluding the ND (No derivative) condition will
Paid science and technology content, Commercial publisher academic journal papers	• Discovering free data sources for non-profit purposes • Purchase paid content licenses • In case of paid data sources, consider quality, coverage, influence
Social Media	• Science and technology content production / service organizations actively disclose the up-to-date information to promote content • Some major social media ResearchGate, Twitter also recommends the use of data for research purposes, so it appears to be highly useful in terms of information reliability

8

4. Results (4)

❑ Guidelines for securing data re-usability

- Copyright guidelines
  - By differentiating from the original information source, work must be classified as an independent work rather than a derivative work, thereby avoiding a situation of copyright infringement.

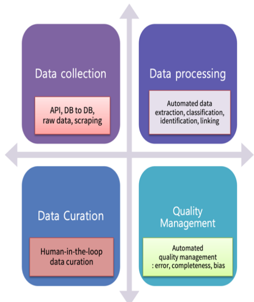
	High risk of copyright infringement	Low risk of copyright infringement
structure	• Maintain the basic structure of the original text	• Differentiating the overall structure from the original excerpt article
Quotation method	• Excerpted directly from the cited article	• Briefly expressed in the author's own way while understanding the original text of the article • translate focusing on the purpose of conveying objective facts, but do not extract the contents of the original text as is, but reorganize it by editing and summarizing words scattered throughout the entire original text.
Quotation amount	• Quoting significant excerpts from the article	• Excerpt from a relatively small portion of the article
etc	-	• quotes at a level that ensures they do not displace demand for the article by affecting the market value of the original text

- Creative Commons (CC) License
  - In order to comply with these conditions, only information whose CC license of the original information source is CC BY, CC BY-NC, CC BY-SA, or CC BY-NC-SA without ND conditions can be used.

9

4. Results (5)

❑ Content Management with AI automation : Technology research



Content collection technology
- Web Crawling, Web Scraping : Apache Nutch, Selenium, NodeJS

Content processing technology
- Extraction of metadata, object : GROBID, Apache Tika, ABBYY Flexi Capture, LayoutParser, PDFMiner

Content generation technology
- Content summaries, explanation generation with generative AI models : GPT, BART, T5, KoBART

Content quality management technology
- Automated quality management : completeness, spelling and grammar errors, AI fairness diagnosis

10

4. Results (6)

❑ Content Management with AI automation : Process-technology diagram

Data collection and data pool construction → **Data processing and creation** → **Data demand survey & analysis** → **Service provision**

Data discovery & collection
- Publisher
- OA repository
- Content provider

Data ingestion
- API or dataset
- web scraping

Metadata extraction
- Automatic extraction of document metadata

Entity extraction and identification
- Tables, figures
- Extract document entities, identification, and linking

Classification
- Apply science and technology standard classification
- Multi-label document classification of convergence studies

Automatic generation (Data creation & management)
- BART/T5 algorithm
- Quality assessment and bias mitigation

User survey
- Survey, interview, email, chatbot, in-person, etc.

Current issue research
- Government Policy, society current issue monitoring
- Discover topics through research proposals, research procurement RFP, etc.

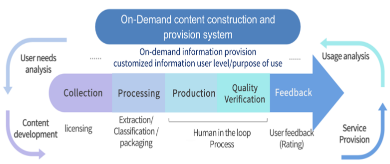
Automatic generation (Human readable documents)
- Publishing documents on current issues
- Human-in-the-Loop

Real-time service
- Automatic Q&A for collected documents (Chatbot, etc.)

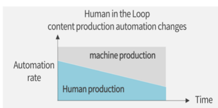
11

4. Results (7)

❑ On-demand content construction and provision system design



Data collection/ processing	• Minimize manual processes by applying collection / processing AI technology
Content production	• Content production using the human-in-the-loop method • Content verification for automatic production
User needs/ Feedback	• Analysis of user information use • User feedback (Content evaluation/rating)



12

5. Conclusion

- It is needed to expand data spectrum and strengthen up-to-date information provision using AI technology to support 4th generation researcher environments
- It is needed to digital transformation of data collection, construction, and management system
- For making those in real, it is required to global cooperation of data management/service institutions, technology corporates, government and NGOs in AI era.

13

Acknowledgement

This research was supported by Korea Institute of Science and Technology Information (KISTI) - Construction on Intelligent SciTech Information Curation (K-23-L01-C01-S01)