

01 Background and Motivation

The Korea Institute of Science and Technology Information (KISTI) has long served as a national institution for collecting, managing, and providing scientific and technical information. As the AI era advances, the role of information curation is expanding beyond document collection and metadata management.

Scientific and technical information now needs to be transformed into structured, linked, reusable, and trustworthy data that AI systems can use for learning, retrieval, reasoning, and evaluation. In this context, KISTI's Data Curation Center is redefining its role from a provider of human-readable information services to a producer of AI-ready scientific data assets.

This transition is especially important for institutions that manage grey literature and other scientific resources, because their holdings contain rich domain knowledge that can be reused as high-quality AI training and evaluation data when properly structured and validated.

02 DiCuON — KISTI's AI-Based Data Curation System

DiCuON is KISTI's integrated data curation system for transforming scientific and technical documents into structured data assets. It applies six AI-based curation technologies to papers, R&D reports, patents, and other scientific resources. Rather than treating documents only as files to be stored and searched, DiCuON enables scientific documents to be decomposed into machine-readable components that can support AI development and data-driven research services. It also includes a DiCuON Labeling Tool that annotates key document regions with color-coded bounding boxes, enabling scientific papers to be converted into structured AI-ready data.

04 Example: One Scientific Paper → Multiple AI-Ready Data Components

Annotation note : The DiCuON Labeling Tool visualizes key document regions with color-coded bounding boxes and converts them into structured AI-ready data.

The image displays six sample document pages from IEEE Access, illustrating the output of the DiCuON Labeling Tool. Each page is annotated with color-coded bounding boxes that identify key document regions, which are then converted into structured AI-ready data components. The components are labeled as follows:

- ① Paper Metadata — CuONEX / CuONid**: Includes title, author names, affiliations, abstract, keywords, DOI, and journal info.
- ② Section Titles & Paragraphs — CuONEX / CuONOTOC / CuONClass**: Classifies S&T content by subject area and document type.
- ③ Tables & Figures — CuONTableCell / CuONOCR**: Recognizes tables, figures, and scientific notations in S&T documents.
- ④ Formulas & Scientific Notations — CuONOCR**: Recognizes formulas, symbols, and scientific notations in S&T documents.
- ⑤ Reference Metadata — CuONEX / CuONid**: Generates hierarchical section structures from S&T documents.
- ⑥ Table Cell Structuring**: Structures tables through cell-level recognition and segmentation.

05 DiCuON's Six AI Curation Technologies

CuONEX	CuONClass	CuONid	CuONOCR	CuONOTOC	CuONTableCell
Information Extraction Detects and extracts bibliographic and content regions in articles and reports. → Title, author names, affiliations, abstract, keywords, DOI, journal info	Document & Subject Classification Classifies S&T content by subject area and document type. → S&T subject categories and document-level classification labels	Entity Identification Identifies and disambiguates authors and institutions in articles, reports, and patents. → Author name disambiguation, institution identification, entity linking	Scientific OCR Recognizes formulas, symbols, and scientific notations in S&T documents. → Formulas, mathematical symbols, tables, figures, and scientific notations	Table of Contents Generation Generates hierarchical section structures from S&T documents. → Hierarchical section structure and table-of-contents data	Table Cell Structuring Structures tables through cell-level recognition and segmentation. → Table detection, cell segmentation, and cell-level structured data

06 AI-Ready Dataset Portfolio (as of April 2026)

Through the DiCuON pipeline, KISTI has built and currently manages approximately 253.4 million structured, machine-readable AI-ready data records:

4.14M records Article Information Extraction	2.15M records Document Layout & OCR	20.12M records Document Understanding & Classification
2.40M records Full-text Papers	217.24M records Bibliographic Data	7.35M records Author & Institution IDs

TOTAL 253.4M AI-ready data records

07 Conclusion and Significance

► KISTI's experience demonstrates that institutions managing grey literature and scientific information resources can serve not only as content providers, but also as producers of structured AI-ready data assets.

► The role of curation is expanding from preservation and access toward AI training data production, enabling scientific information institutions to create new value from existing document collections.

► By transforming papers, reports, patents, and other scientific resources into structured and reusable components, KISTI can support AI model training, retrieval-augmented generation, scientific data services, and future research intelligence applications.

► The DiCuON pipeline provides a scalable model for converting scientific documents into trustworthy AI-ready data assets, while maintaining the domain-specific structure and quality required for scientific and technical information.

08 Future Work

► Design and implement a KISTI AI-Ready Data schema reflecting S&T information characteristics, including DOI, document type, formula presence, evidence tracing, and S&T classification fields.

► Develop an automated AI-ready data export function within the DiCuON pipeline, enabling curated document components to be reused for both information services and AI training data construction.

► Explore Croissant-compatible metadata descriptions for KISTI's AI-ready datasets to improve findability, interoperability, and reuse in AI research and development contexts.

► Develop a controlled AI training data catalogue and API-based access model, subject to data rights, licensing conditions, and institutional access policies.

AI-Ready Data	Grey Literature	Information Curation	Scientific Data	DiCuON
AI Training Data	Data Quality	FAIR Principles	Document Understanding	Knowledge Infrastructure