

<https://doi.org/10.5281/zenodo.17525311>

Denkov Ivan,
MS,

V.M. Glushkov Institute of Cybernetics,
National Academy of Sciences of Ukraine,
ORCID: 0009-0007-7989-2169; e-mail: havoc85@gmail.com

Kolomiyets Oleksandr,
PhD, senior researcher,

V.M. Glushkov Institute of Cybernetics,
National Academy of Sciences of Ukraine,
ORCID: 0009-0003-2542-8388; e-mail: skolomon@gmail.com

Tulchinsky Vadim,

Dr.Sc., senior researcher,
head of Programming Automation Department,
V.M. Glushkov Institute of Cybernetics,
National Academy of Sciences of Ukraine,
ORCID: 0000-0002-0280-223X; e-mail: dep145@gmail.com

ON AUTOMATED METADATA GENERATION FOR PREPRINTS AND RESEARCH PAPERS

ABSTRACT. Large Language Models (LLMs) are capable of high-quality extraction of key descriptive components (metadata) from a text of scientific work, including its title, authors, the author affiliations, publication date, keywords, and abstract. The missing components are generated from the text if possible. The paper's language is detected, and bilingual metadata are obtained with translating between English and Ukrainian where needed. Using LLM GPT-4o mini, a specialized webservice for automatic metadata extraction from research texts has been developed. It processes PDF documents and generates metadata in the international format of Dublin Core. The webservice automates and simplifies the process of submitting publications to the Archive of Preprints of the National Academy of Science (NAS) of Ukraine. The same approach and tool can be used by repositories of research texts for the rapid import of e-libraries and institutional collections that do not contain metadata in right form. It can also be recommended for journal submission automation.

KEYWORDS: open science, metadata, open access, e-libraries, preprints, research texts, feature extraction, large language models, LLM.

INTRODUCTION. Open science is among the top priorities for NAS of Ukraine on its way to European Research Area (ERA) [1]. Since 2023, NAS of Ukraine has developed the Open Science Infrastructure [2] that addresses regulatory, organizational, software and technical challenges of implementing open science on a large scale. NAS of Ukraine consists of more than 200 research institutions. According to the plan, they are to adopt open access to research data and to transform their scientific journals for open access in 2 years. Open datasets and preprints have been added to the list of project implementation metrics. One of the most complex tasks in the transition is training scientists and helping them become accustomed to the new rules. The simplicity of the new actions and tools is crucial for a successful transition.

The institutional archive of preprints called Arxiv Academy [3] is an important component of the Open Science Infrastructure of NAS of Ukraine. In open science, preprints complement open data by protecting the authorship of ideas linked to that data when a paper is submitted to a journal. Of course, preprints can also be published as separate manuscripts for discussion within the research community before being submitted for peer review. Arxiv Academy is a fully functional preprint repository that issues Digital Ob-

ject Identifiers (DOI) to all published preprints and integrates their metadata into international search sites such as OpenDOAR, OpenAIRE, BASE, and CORE. The platform's user interface can be presented in Ukrainian or English with preprints accepted in any of the two languages. Arxiv Academy is based on the PKP OPS [4] but is significantly customized to meet the needs and rules of NAS of Ukraine. Automated generation of bilingual metadata — including title, authors, author affiliations, publication date, keywords, and abstract — became a new customization of the archive software, based on feature extraction and full-text processing.

MATERIALS AND METHODS. The introduction of LLMs — beginning with the prominent OpenAI ChatGPT — is ushering in a new era of human-computer interaction and the automation of routine operations. The new tools were quickly adapted for science and are now used in popular systems such as Google NotebookLM, Paperpal Generative AI, Jenni AI Research Assistant, ThesisAI Academic Writing, Avidnote AI for Research, and SciSpace AI Research Agent. We apply them to automatically complete the input forms for preprint submission. Anticipating more extensive future use, we implemented the LLM interface as a webservice, which is shown in Fig. (1).

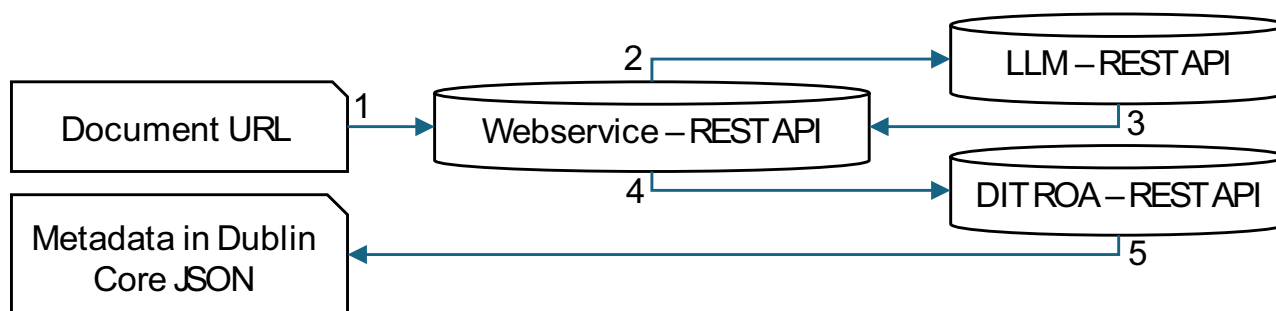


Fig. 1. Data stream in the Dublin Core metadata generation webservice

The webserver features a REST API program interface that handles requests and responses in JSON format. The main request contains a local or remote link (URL) to a PDF document. The webservice extracts the document content and generates an LLM request based on it. The obtained response (already in the Dublin Core JSON format [5]) is verified and supplemented using DIT ROA [6]. DIT ROA is applied to fix names and attributes of authors and institutions.

RESULTS. The Arxiv Academy's Publication Submission Wizard has been changed. The first page is now the preprint import form, where the users can choose to have the LLM process their submitted text or opt out. In our tests, the LLM performed with high quality and manual changes were rarely needed in the subsequent steps. However, the operation requires an advanced enough model. We've selected the GPT-4o mini. Usually, the result is returned in several seconds, but sometimes the processing takes up to 30 seconds. The most frequent issues were missing data (such as DOIs in older papers), and unusual errors in determining the text's language.

CONCLUSION. The developed webservice will be further improved and implemented within the Open Science Infrastructure of NAS of Ukraine for wider use. Its target applications include the automation of the metadata assignment for old collections and outdated e-libraries, where full texts of research papers were available for open access but lacked metadata or standard protocol OAI PMH [5] for their acquisition. The webservice support can also be included in the Universal Publishing Platform [1], which is another part of the Open Science Infrastructure. Using an LLM to test the correctness of metadata assigned to research texts is planned for further development.

FUNDING: This work was financially supported by the Targeted Scientific and Technical Project of NAS of Ukraine “Implementation and support of open science in the institutions of the NAS of Ukraine (OPENS2)” for 2025-2026.

REFERENCES

1. Zagorodny, A. Khimich, O. M., Andon, P. I. (2025). Implementation of European principles of open science in the National Academy of Sciences of Ukraine. *Visnyk of Nat. Acad. of Sc. of Ukraine*, 1, 11-33. <https://doi.org/10.15407/vsn2025.01.011> [in Ukrainian].
2. Tulchinsky, V., Svistunov, S., & Khimich, O. (2023). On the Distributed Open Research Infrastructure of the National Academy of Sciences of Ukraine. *Open Science and Innovation in Ukraine*, 117-120. <https://doi.org/10.2174/9789815256956124010034> [in Ukrainian].
3. Kolomiyets, O., Lavreniuk, S., & Pustovoi, M. (2023). Development and Deployment of the “Arxiv Academy” Archive of Preprints for Ukrainian Scientists. *Open Science and Innovation in Ukraine* 2023, 43-46. <https://doi.org/10.2174/9789815256956124010016> [in Ukrainian].
4. Public Knowledge Project (PKP). <https://pkp.sfu.ca/ops>
5. Dublin Core Initiative. (2026). <https://www.dublincore.org/>
6. Khimich, O. Pustovoi, M. M., & Ivlichev, V. P. (2018). A framework for the creation of distributed information technology to support the scholarly research and organizational activities of the NAS of Ukraine. *Sci. Innov.* 14(1), 47–59. <https://doi.org/10.15407/scine14.01.047>

Денков Іван,
магістр,

Інститут кібернетики імені В.Глушкова НАН України,
ORCID: 0009-0007-7989-2169; e-mail: havoc85@gmail.com

Коломієць Олександр,

кандидат технічних наук, старший науковий співробітник,
Інститут кібернетики імені В.Глушкова НАН України,
ORCID: 0009-0003-2542-8388; e-mail: skolomon@gmail.com

Тулчинський Вадим,

доктор економічних наук, старший науковий співробітник,
Інститут кібернетики імені В.Глушкова НАН України,
ORCID: 0000-0002-0280-223X; e-mail: dep145@gmail.com

ПРО АВТОМАТИЗОВАНУ ГЕНЕРАЦІЮ МЕТАДАНИХ ДЛЯ ПРЕПРИНТІВ ТА ДОСЛІДНИЦЬКИХ СТАТЕЙ

АНОТАЦІЯ. Великі мовні моделі (LLM) здатні високоякісно витягувати ключові описові компоненти (метадані) з тексту наукової роботи, включаючи її назву, авторів, афіліацію авторів, дату публікації, ключові слова та анотацію. Відсутні компоненти генеруються з тексту, якщо це можливо. Визначається мова статті, та отримують двомовні метадані з перекладом англійською та українською мовами, де це необхідно. За допомогою LLM GPT-4o міні розроблено спеціалізований веб-сервіс для автоматичного видобутку метаданих з дослідницьких текстів. Він обробляє PDF-документи та генерує метадані в міжнародному форматі Dublin Core. Веб-сервіс автоматизує та спрощує процес подання публікацій до Архіву препринтів Національної академії наук (НАН) України. Такий самий підхід та інструмент можуть використовуватися репозитаріями дослідницьких текстів для швидкого імпорту електронних бібліотек та інституційних колекцій, які не містять метаданих у належному вигляді. Його також можна рекомендувати для автоматизації подання до журналів.

КЛЮЧОВІ СЛОВА: відкрита наука, метадані, відкритий доступ, електронні бібліотеки, препринти, дослідницькі тексти, вилучення ознак, моделі великих мов, LLM.