**Денков Іван Дем’янович,**

ORCID <https://orcid.org/0009-0007-7989-2169>,

аспірант,

[відділ автоматизації програмування](https://incyb.kiev.ua/podrzdeleniya/viddilennya-matematichnoi-kibernetiki-ta-sistemnogo-analizu/viddil-avtomatizaciyi-programuvannya),

Інститут кібернетики імені В. М. Глушкова,

Національна академія наук України,

Київ, Україна

e-mail: havoc85@gmail.com

**Отримання метаданих з наукових робіт за допомогою штучного інтелекту**

Розглянуто можливість автоматизації отримання метаданих із наукових праць за допомогою штучного інтелекту. Створено вебсервіс на Python з використанням GPT-4o mini, що екстрагує з PDF-файлів структуровані метадані у форматі JSON. Наголошено на необхідності валідації результатів людиною.

*Ключові слова:* наукові метадані, відкрита наука, штучний інтелект, великі мовні моделі, репозитарії даних

Для втілення в життя принципів відкритої науки створюється відповідна інфраструктура: репозитарії відкритих даних, харвестери, електронні бібліотеки та інші службові елементи. Подібна діяльність вимагає структуризації даних. Зокрема, треба отримати метадані наукових робіт і привести їх до певних стандартів.

Зараз в Україні ця робота виконується вручну співробітниками наукових закладів з усіма відповідними недоліками: низька швидкість виконання, помилки внаслідок людського фактору тощо. Отже постає питання пришвидшення і оптимізації процесу. Нами було досліджено можливість задіяти штучний інтелект (ШІ) у процесі отримання метаданих з наукових праць.

В роботі використали LLM ChatGPT, модель GPT-4o mini. Створено вебсервіс за допомогою мови програмування Python на основі фреймворку Flask. Механізм роботи такий: на сервер відправляється POST-запит з JSON-даними, які містять посилання на наукову роботу у форматі PDF; текст файлу подаємо на вхід ШІ разом зі сформульованою спеціальним чином підказкою; на виході – JSON-файл з метаданими англійською та українською, форматованими згідно стандартів Дублінського ядра. За умови наявності з наукових праць планували отримати наступні метадані:

* Автор(и)
* Назва статті
* Анотація
* Ключові слова
* Мова статті
* Періодичне видання, де опубліковано статтю
* Рік та номер видання, де опубліковано статтю
* Видавець
* Посилання на вихідний PDF-файл
* Бібліографічне посилання в заданому форматі
* Код ISSN періодичного видання
* УДК
* DOI

Для перевірки роботи сервісу використовувалися матеріали, що відносяться до різних галузей науки з різних джерел (електронні бібліотеки, періодичні видання, архів препринтів НАН України). Приклад англомовної частини результату наведено нижче; атрибути метаданих відсортовано за алфавітом:

*{*

 *"dc.contributor.author": "Pavlo Maslianko, Kate Pavlovska",*

 *"dc.date.issued": "2024, No. 4",*

 *"dc.description.abstract": "The article proposes a conceptual model and NLP system 'Text to image' based on the methodology of system engineering of Data Science systems, architecture, and software of the image generation system based on the latent diffusion model. It is proposed to improve the basic architecture of the latent diffusion model by using a diffusion transformer. It is found that unlike approaches based on U-Net architecture, DiTs work with latent patches, providing better scalability and increased performance. The purpose of the work is to develop a scientifically based conceptual model and system for transforming text descriptions into images, based on the methodology of system engineering, modern methods of deep learning and business profile of Erikson – Penker. Verification and validation of the developed NLP system 'Text to image' for converting text data into images was carried out. The generation results demonstrate the exact reproduction of key elements, which indicates the high quality of the correspondence between the image and the text description. As a result of a comparative analysis of the performance of the models, it was determined that the TransformerLD system, although inferior to the Stable Diffusion and DALL-E 2 models in terms of FID and IS, still remains competitive.",*

 *"dc.identifier": "http://cctech.org.ua/images/docs/Articles/2024/paper\_24\_4\_9.pdf",*

 *"dc.identifier.citation": "Maslianko, P., & Pavlovska, K. (2024). Conceptual Model and NLP-System 'Text to Image'. Cybernetics and Computer Technologies, 4.",*

 *"dc.identifier.issn": "2707-4501",*

 *"dc.identifier.udc": "519.688",*

 *"dc.identifier.uri": "10.34229/2707-451X.24.4.9",*

 *"dc.language.iso": "en",*

 *"dc.publisher": "Igor Sikorsky Kyiv Polytechnic Institute",*

 *"dc.relation.ispartof": "Cybernetics and Computer Technologies",*

 *"dc.subject": "system engineering, Data Science, NLP-systems 'Text to image'",*

 *"dc.title": "Conceptual Model and NLP-System 'Text to Image'"*

 *}*

Аналіз результатів вказує, що ШІ може отримати метадані з наукової роботи, а деякі згенерувати за потреби. Інколи в його роботі можливі помилки. Так, наприклад, якщо в статті є анотація двома мовами, то в англійській частині JSON пишеться, що мова статті англійська, а в українській – українська. Також, якщо якісь дані в тексті відсутні, ШІ знайти їх не може. Отже, співробітник має проводити ревізію результатів роботи сервісу і коригувати помилки. Підказка (набір атрибутів метаданих) для ШІ може змінюватись відповідно до запитів різних закладів.

Таким чином, вебсервіс за умови мінімальної участі людини може стати повноцінною частиною інфраструктури відкритої науки, даючи можливість електронним бібліотекам, видавництвам, сайтам дослідницьких установ перетворити свої колекції наукових текстів на повноцінні репозитарії, здатні передавати метадані в харвестери з подальшою інтеграцією в міжнародні пошукові системи відкритої науки.

**Ivan Denkov,**

ORCID <https://orcid.org/0009-0007-7989-2169>,

Postgraduate,

Programming Automation Department,

V. M. Glushkov Institute of Cybernetics,

National Academy of Sciences of Ukraine,

Kyiv, Ukraine

e-mail: havoc85@gmail.com

**Extracting metadata from scientific papers using artificial intelligence**

The article considers the possibility of automating the extraction of metadata from scientific papers using AI. A web service in Python was created using GPT-4o mini, which extracts structured metadata in JSON format from PDF files. The need for human validation of the results was emphasized.

*Keywords:* scientific metadata, open science, artificial intelligence, large language models, data repositories