BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 7: e112428
doi: 10.3897/biss.7.112428

OPEN ACCESS

Conference Abstract

"Publish First": A Rapid, GPT-4 Based Digitisation System for Small Institutes with Minimal Resources

Rukaya Johaadien‡, Michal Torma‡

‡ Natural History Museum, University of Oslo, Oslo, Norway

Corresponding author: Rukaya Johaadien (rukayaj@gmail.com), Michal Torma (miso.torma@gmail.com)

Received: 08 Sep 2023 | Published: 11 Sep 2023

Citation: Johaadien R, Torma M (2023) "Publish First": A Rapid, GPT-4 Based Digitisation System for Small Institutes with Minimal Resources. Biodiversity Information Science and Standards 7: e112428. https://doi.org/10.3897/biss.7.112428

Abstract

We present a streamlined technical solution ("Publish First") designed to assist smaller, resource-constrained herbaria in rapidly publishing their specimens to the Global Biodiversity Information Facility (GBIF).

Specimen data from smaller herbaria, particularly those in biodiversity-rich regions of the world, provide a valuable and often unique contribution to the global pool of biodiversity knowledge (Marsico et al. 2020). However, these institutions often face challenges not applicable to larger herbaria, including a lack of staff with technical skills, limited staff hours for digitization work, inadequate financial resources for specialized scanning equipment, cameras, lights, and imaging stands, limited (or no) access to computers and collection management software, and unreliable internet connections. Data-scarce and biodiversity rich countries are also often linguistically diverse (Gorenflo et al. 2012), and staff may not have English skills, which means pre-existing online data publication resources and guides are of limited use.

The "Publish First" method we are trialing, addresses several of these issues: it drastically simplifies the publication process so technical skills are not necessary; it minimizes administrative tasks saving time; it uses simple, cheap and easily available hardware; it

© Johaadien R, Torma M. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

does not require any specialized software; and the process is so simple that there is little to no need for any written instructions.

"Publish first" requires staff to attach QR code labels containing identifiers to herbarium specimen sheets, scan these sheets using a document scanner costing around €300, then drag and drop these files to an S3 bucket (a cloud container that specialises in storing files). Subsequently, these images are automatically processed through an Optical Character Recognition (OCR) service to extract text, which is then passed on to OpenAI's Generative Pre-Transformer 4 (GPT-4) Application Programming Interface (API), for standardization. The standardized data is integrated into a Darwin Core Archive file that is automatically published through GBIF's Integrated Publishing Toolkit (IPT) (GBIF 2021).

The most technically challenging aspect of this project has been the standardization of OCR data to Darwin Core using the GPT-4 API, particularly in crafting precise prompts to address the inherent inconsistency and lack of reliability in these Large Language Models (LLMs). Despite this, GPT-4 outperformed our manual scraping efforts. Our choice of GPT-4 as a model was a naive one: we implemented the workflow on some pre-digitized specimens from previously published Norwegian collections, compared the published data on GBIF with GPT-4's Darwin Core standardized output, and found the results satisfactory. Moving forward, we plan to undertake more rigorous additional research to compare the effectiveness and cost-efficiency of different LLMs as Darwin Core standardization engines. We are also particularly interested in exploring the new "function calling" feature added to the GPT-4 API, as it promises to allow us to retrieve standardized data in a more consistent and structured format.

This workflow is currently under trial in Tajikistan, and may possibly be used in Uzbekistan, Armenia and Italy in the near future.

Keywords

AI, LLMs, digitisation

Presenting author

Rukaya Johaadien

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- GBIF (2021) Darwin Core Archives – How-to Guide, version 2.2. Copenhagen: GBIF Secretariat. URL: https://ipt.gbif.org/manual/en/ipt/2.5/dwca-guide
- Gorenflo LJ, Romaine S, Mittermeier RA, Walker-Painemilla K (2012) Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. Proceedings of the National Academy of Sciences of the United States of America 109 (21): 8032-7. https://doi.org/10.1073/pnas.1117511109
- Marsico T, Krimmel E, Carter JR, Gillespie E, Lowe P, McCauley R, Morris A, Nelson G, Smith M, Soteropoulos D, Monfils A (2020) Small herbaria contribute unique biogeographic records to county, locality, and temporal scales. American Journal of Botany 107 (11): 1577-1587. https://doi.org/10.1002/ajb2.1563