

Final Technical Report

A Proof-of-Concept Metadata Catalogue and Online Portal of Earth Observation Datasets for Human Health Research in Exposomics

July 2024

Maged N. Kamel Boulos^{1,*}, with Keumseok Koh², Gang Zheng¹, Hongsheng Zhang¹, Muralikrishna V. Iyyanki³, Bosco Bwambale⁴ and Ashraf Dewan⁵

¹ Project P.I., University of Lisbon, 1649-028 Lisbon, Portugal

² Dept. of Geography, Faculty of Social Sciences, The University of Hong Kong, Hong Kong 999077, China

³ JN Technological University Hyderabad, Hyderabad, 500008, India

⁴ Mountains of The Moon University, P. O. Box 837, Fort-Portal, Uganda

⁵ School of Earth & Planetary Sciences, Curtin University, Perth, WA 6845, Australia

Executive Summary: This is a report about the proof-of-concept research we carried out during 2023 and the first quarter of 2024 under an International Society for Photogrammetry and Remote Sensing (ISPRS)-funded project to develop and disseminate a metadata catalogue of Earth observation data sources/products and types that are relevant to human health research in exposomics, as a free service to interested researchers worldwide. The proof-of-concept catalogue was informed by input from existing research literature on the subject (desk research), as well as online communications with, and relevant research publications collected from, a small panel of select experts from the academia in three countries (China, UK and USA). It has 90 metadata records of relevant Earth observation datasets and associated health-focused research publications. The project's online portal offers a searchable version of the catalogue featuring a number of search modes and filtering options, in addition to a summary of the project's details and methods, including results of our experimentation with a LLM (large language model)-powered interface in 2024, as well as a list of the project's open access outputs and publications. It is hoped future, more comprehensive versions of this service will enable more researchers and studies to discover and use remote sensing data about population-level exposures to disease determinants (exposomic determinants of disease) in combination with other relevant data to reveal fresh insights that could improve our understanding of relevant diseases, and hence contribute to the development of better-optimized prevention and management plans to tackle them.

Keywords: exposomics; human exposome; Earth observation data; remote sensing; data sources; metadata catalogue; public health informatics

Introduction: Vision and Scope

In this International Society for Photogrammetry and Remote Sensing (ISPRS)-funded project under ISPRS Scientific Initiatives 2023 (SI2023 Awards): Geospatial database for exposomics - <https://www.isprs.org/news/announcements/details.aspx?ID=314>, we have proof concepted and disseminated initial development concepts about a much-needed metadata catalogue of Earth observation data sources/products and types that are relevant to human health research. The proof-of-concept searchable catalogue takes the form of a dedicated Geoexposomics Web portal that is provided as a free service to interested parties worldwide at <https://geoexposomics.org/>.

This small project has been conceived as early preparation work/initial exploratory research phase towards future, much more comprehensive versions of the Geoexposomics portal with the ultimate goal of helping more researchers and studies to discover and use remote sensing data about population-level exposures to disease determinants in the years to come. Such data on the exposomic determinants of disease in combination with other relevant data can reveal fresh insights that could improve our understanding of relevant disease conditions, and hence contribute to the development of better-optimized prevention and management plans to tackle them.

The reader should note that this work is strictly a proof of concept (PoC). Whilst perfectly fit for the purpose they are intended to serve, PoCs have inherent (by design) limitations compared with more developed prototypes and pilots. The latter serve different purposes and would have required much larger budgets and resources to execute than what we had at our disposal for this mini-project. For a detailed comparison of PoCs, prototypes and pilots, please consult: The London School of Economics and Political Science (LSE). Use of Proof of Concepts, Prototypes and Pilots. <https://info.lse.ac.uk/staff/services/Policies-and-procedures/Assets/Documents/proConProPil.pdf> (accessed on 29 July 2024).

Collecting Inputs

The proof-of-concept catalogue was informed by extensive input from existing research literature on the subject (desk research) as well as online communications with, and relevant research publications collected from, a small panel of select 'stakeholder representatives' from the academia in three countries (China, UK and USA) to guarantee some very basic form of user engagement/involvement in the project from its very early stages, which is always highly desirable. Stakeholder representatives provided us with valuable insights and intelligence from their own research experience about relevant remote sensing applications and available products in public health, in addition to important feedback on our catalogue (while we were developing it over two iterations), which helped us shape and improve/refine its concepts to better serve their needs and those of the wider research community they represent.

Dataset

The research team compiled the catalogue based on all collected input in the form of an **Open Access**  **Excel workbook that can be downloaded from CERN's Zenodo Repository at <https://doi.org/10.5281/zenodo.8186444>**

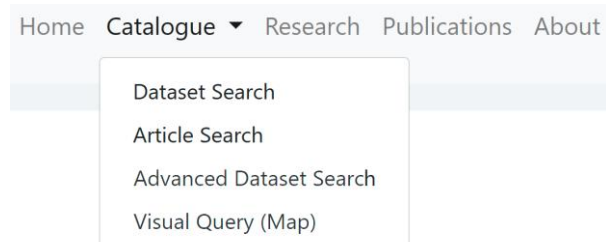
The catalogue currently (at the time of writing this report) has 90 metadata records of relevant Earth observation datasets and associated health-focused research publications. Dataset and research publication records are stored in two separate but fully cross-referenced Excel worksheets/database tables. For each catalogued record, the following metadata elements were compiled: dataset name, provider, available year(s)/update frequency, cost (licence), covered geographical areas, resolution used, Web address (URL), related research publications, and the health conditions/application areas that featured in those publications.

Web Implementation

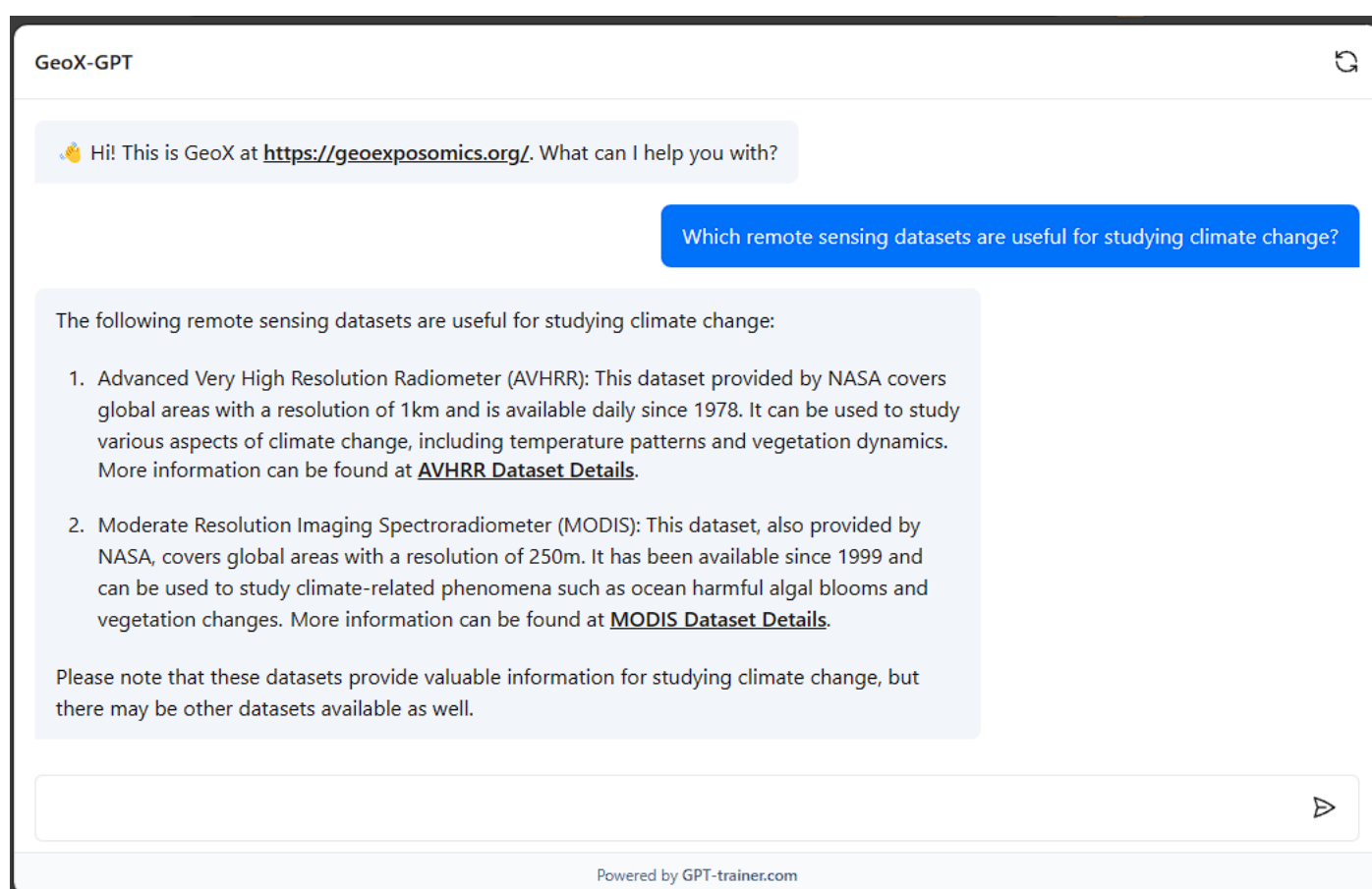
Our Web development team then converted the Excel workbook into an online database and developed the project's portal in MySQL, an open-source relational database management system, Python, a high-level, general-purpose programming language, and Django, an open source, Python-based Web framework. The portal features an access-protected 'Catalogue Management Interface' where the catalogue curator (project team) can modify existing catalogue records or add new ones as necessary.

An online searchable version of the catalogue can be accessed by clicking the 'Catalogue' menu on the project portal's navigation bar at <https://geoexposomics.org/>. The interface was designed with both desktop and mobile devices in mind. It allows users to search and filter dataset records by health conditions/application areas, cost, covered area(s), provider, or a combination of one or more of these elements. Users can also filter the associated research article records by health conditions/application areas and/or dataset name. Soundex search is supported (added during the second

iteration of the interface) to take care of minor typos and spelling variants in users' queries (further details below).



Web interface development also involved a latter more experimental phase in 2024 to conceive GeoX-GPT, an early preview of a possible future ChatGPT-like interface for our metadata catalogue powered by Retrieval-Augmented Generation (RAG), whereby our catalogue became an external knowledgebase to GPT-4 (a Large Language Model), in order to provide more relevant and better cross-linked search results in response to user's prompts.



Screenshot of GeoX-GPT. This experimental RAG implementation was developed in GPT-trainer. It features a natural language-based AI-driven conversational user interface to enhance discovery and explorability by allowing the search user interface to more flexibly adapt to users' needs in more natural ways instead of asking users to adapt to and master a rather rigid (preprogrammed) and less forgiving conventional interface.

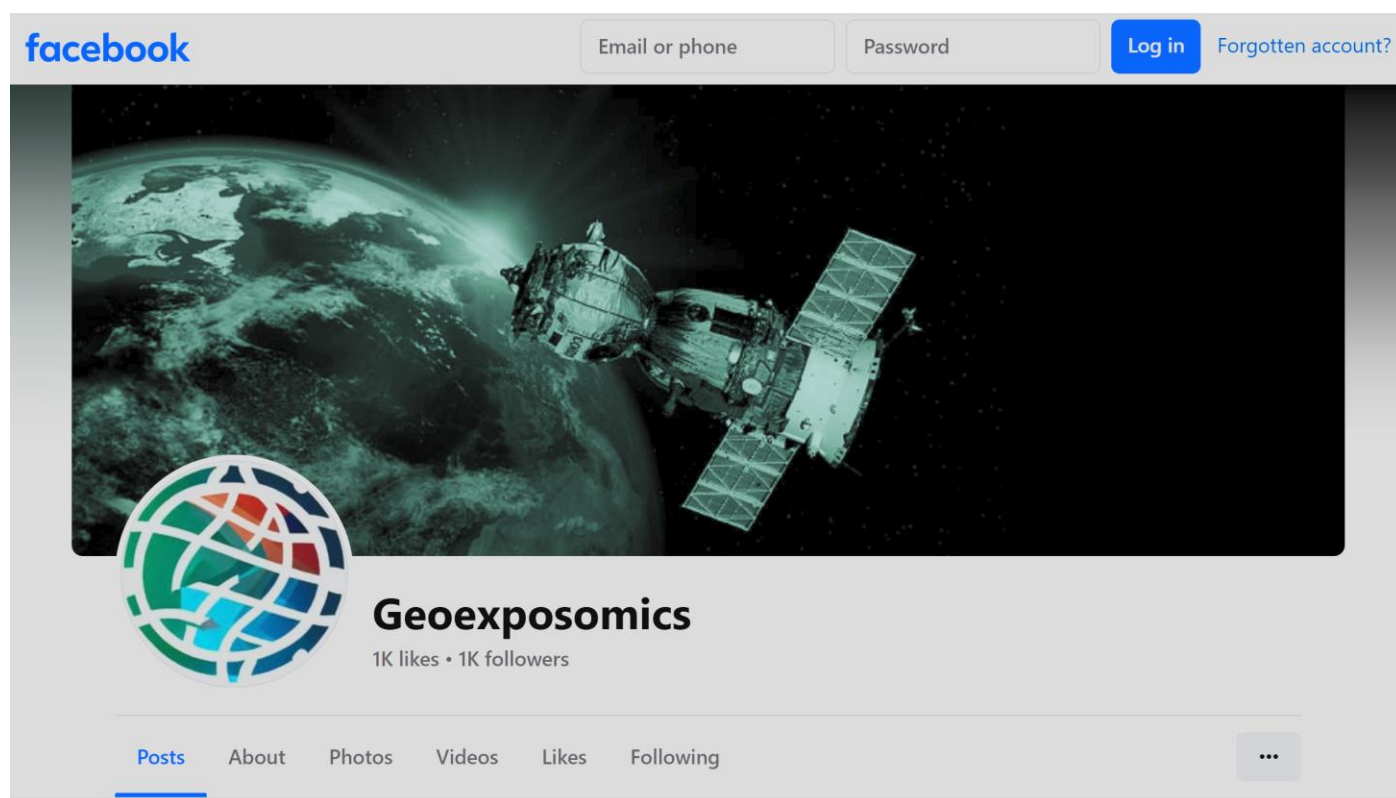
User-Guided Evaluation and Refinement of Initial Concepts

Following the initial launch of the portal and Catalogue in June 2023, we contacted the same group of 'stakeholder representatives' who had responded to our original call and provided their input for the catalogue. The feedback received about the online portal and catalogue was positive and provided valuable suggestions for the second iteration of the interface, namely a suggestion "to implement free text search and ensure it can tolerate different spelling variants (e.g., British and American) and misspellings of health applications and place names" and another one to have a handy "online map to help locating datasets".

The two suggestions were considered and implemented in the second iteration of the portal and catalogue. We implemented an 'Advanced Dataset Search' option using a custom Soundex algorithm (free text search of datasets by health applications or covered areas). As an example of how this algorithm works, searching health conditions for 'haemorrhagic fever' (en-GB spelling), 'hemorrhagic fever' (en-US spelling), or 'hemoragic fever' (typo/misspelling) will all retrieve the correct/same results. Furthermore, during 2024, we conceived GeoX-GPT, an additional GPT-4-powered interface to better handle user questions and interactions with the catalogue content (see details elsewhere in this report). We also added a 'Visual Query (Map)' option allowing users to click location 'hotspots' on a map of the world with country flags and other relevant icons to retrieve corresponding dataset records (datasets by covered areas).

Dissemination and Publication Activities

As of 30 June 2024, the portal received >2500 unique visitors (repeat visits by same IP address are not counted) from more than 40 countries according to anonymous server log statistics generated by AWStats tool using IP geolocation. We also have a companion Facebook page at <https://www.facebook.com/geoexposomics> where relevant news items and updates are regularly shared with a growing community of more than 1000 followers (as of 30 June 2024).



Geoexposomics' virtual community on Facebook.

We also used the ISPRS e-Bulletin (Issue No 5 – 2023) to raise further awareness about the project by having some information and links published under its 'Society News' section: <https://www.isprs.org/news/newsletter/2023-05/index.html#society>.

The following research outputs were formally published:

- **ISPRS Conference Paper:** Kamel Boulos MN, Iyyanki MV, Dewan A, Bwambale B, Koh K. Development of an online portal and metadata catalogue of Earth observation data types, sources and products for human health

research in exposomics (Extended Abstract/Oral Presentation). Presented at *ISPRS Geospatial Week 2023 (Egypt GSW'2023)*, Cairo, Egypt, 2-7 September 2023 (Track: ADP-2: Advanced Data Preparation & Data Management for Geospatial & Remote sensing 2, Monday, 4 Sep 2023, 3:30pm-5:00pm, Pavilion-Hall3, InterContinental Cairo Semiramis). <https://gsw2023.com/> - doi:10.6084/m9.figshare.24845022 [[Extended Abstract \(PDF\)](#)], **Project Presentation:** [[PowerPoint Presentation \(PDF\)](#)] as delivered at GSW'2023.

- **Full Paper** (Peer Reviewed, Open Access): Koh K, Kamel Boulos MN, Zheng G, Zhang H, Iyyanki MV, Bwambale B, Dewan A. A proof-of-concept online metadata catalogue service of Earth observation datasets for human health research in exposomics. *Journal of the Academy of Public Health*. 2025 Jan, 1, pp.: accepted/in press. **Preprint (identical to accepted version):** [[arXiv:2311.08770](https://arxiv.org/abs/2311.08770)]

The following **Acknowledgments** appear at the end of the above paper:

- **Funding:** This research was funded by the International Society for Photogrammetry and Remote Sensing (ISPRS) under ISPRS Scientific Initiatives 2023 (SI2023 Awards).
- **Data Availability Statement:** The proof-of-concept catalogue data covered in this paper are publicly and freely available from doi: 10.5281/zenodo.8186444 and <https://geoexposomics.org/>
- The authors would like to express their gratitude to the panel of stakeholder representatives that were involved in the project.
- **Dataset** (Open Access): Kamel Boulos MN, GeoExposomics.org Project Team. GeoExposomics.org proof-of-concept catalogue as an Excel workbook (Catalogue Main Version: 10 May 2023; Last Updated 25 July 2023) [Data set]. *Zenodo*. 2023. <https://doi.org/10.5281/zenodo.8186444>

Future Directions

A follow-on full project would take the concepts developed in this PoC exercise and prototype then pilot them in a more advanced version of the catalogue and portal guided by a much wider stakeholder consultation. Details of possible future work are in the above-mentioned project presentation and full paper.