Progress with Improving Preservation and Reuse of Scientific Research Data

Donat Agosti
Plazi, Switzerland

Peter Cornwell
Data Futures, Germany

Laurence Benichou
MNHN, France

Jose Gonzalez
CERN, Switzerland

Madeleine Herren
EIB, Switzerland

Patrick Ruch
SIB, Switzerland

Abstract – We report progress towards automatically transforming existing analyses of scientific literature into annotations based on W3C's Web Annotation Data Model (WADM). Case studies are presented from the life sciences, and social sciences and humanities, in which these developments have led to the creation of new unrestricted data services for the research community. We discuss the cross-domain potential of annotation infrastructure for releasing scientific facts reported in research literature from copyright restrictions, and demonstrate the utility of common standards-based preservation and discovery methods in disparate activities. We suggest that scientific treatments of literature using WADM annotation can lead to new mechanisms for access to and reuse of research data, and accelerate convergence with the FAIR Principles.

I. Introduction

Effective reuse and access to existing scientific research outputs has gained importance across research domains during the last decade. In particular, the ability of the Open Science movement to oversail national boundaries has been instrumental in addressing the global COVID pandemic. On one hand it became recognized at the beginning of the 21st century that research undertaken using digital methods was vulnerable to loss through cyclic obsolescence of technical infrastructure on which research data resided, or because of organizational change, or combinations of these factors. New long-term preservation practices were urgently required to stem substantial loss of research investments. Addressing this crisis has involved changing the guidelines issued by funding agencies as well as researchers' practices, together with the development of new preservation technologies and their uptake, but progress in both areas has been slow. On the other hand, improving the reuse of newly-published scientific literature—led by the European Commission—has increased open access to publications resulting from publicly-funded research, though publishers' pay-walls remain significant impediments, and access to and preservation of underlying research data remains poor.

Significantly, these endeavors do not address *reuse of scientific facts reported in existing literature*, which is pivotal for activities such as biodiversity and Earth sciences because historical records report conditions that cannot be replicated: surviving research results are embedded in the literature. We present three case studies in which new standards-based approaches have been adopted to overcome technology obsolescence and copyright restrictions, and so improve reuse of existing scientific literature and the reuse of data generated through historic and current research investments more generally.

II. Enriching Copyright-free Bio-taxonomy Data Derived from Existing Literature

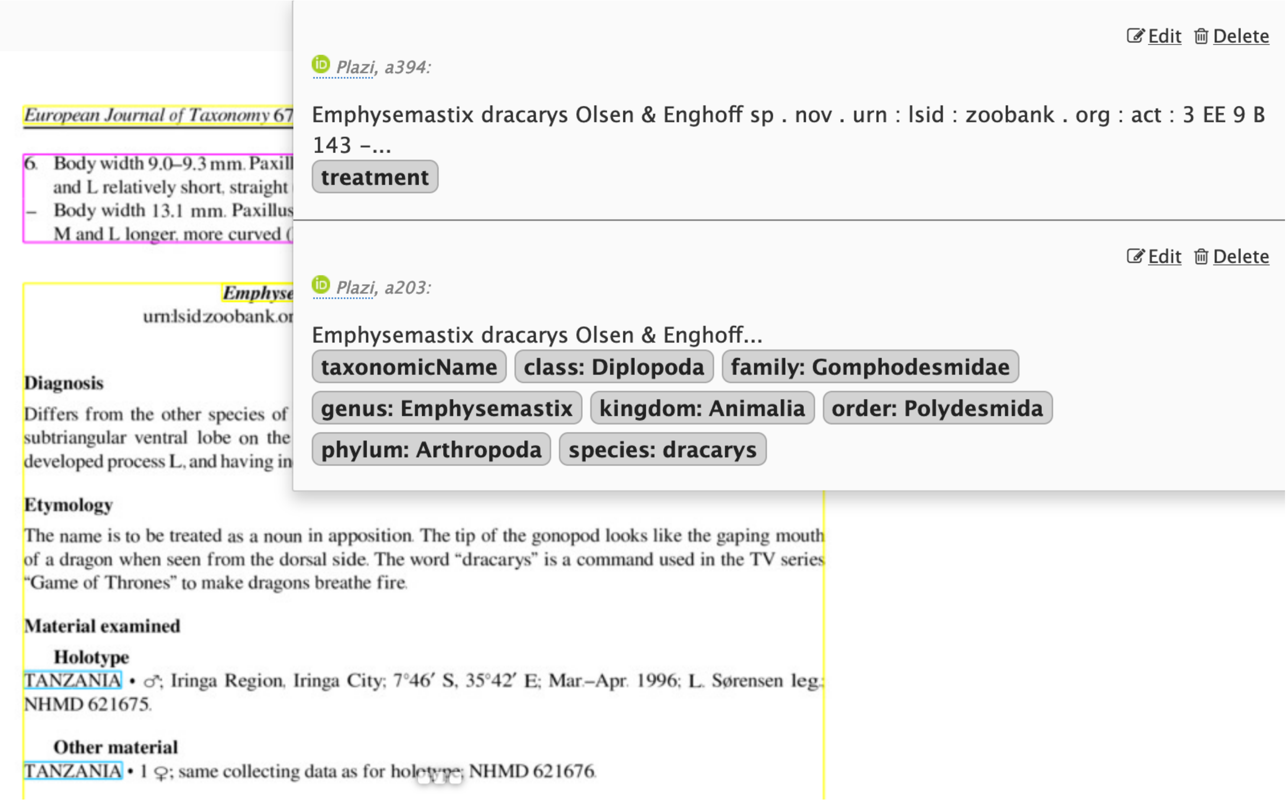
Multiple domain-specific systems have evolved for defining key scientific facts electronically within published literature (named entities, text fragments, tables and illustrations, etc), for the purpose of connecting external metadata so that they can be located independently from (re)reading. There are challenges making such metadata preservable, since it is generally stored using separate digital infrastructure—often with independent maintenance constraints—but robust, standards-based preservation methods have been developed which entered service in 2019 (see case study from the humanities, below). This process of 'annotation', which can be refined based on the scientific domain, augments bulk machine discovery of facts reported in publications and it can potentially return an excerpt interactively—enabling further assessment in original context.

Identifying sections of literature for this purpose is referred to as target definition, and such targets form an integral part of annotation metadata, which is increasingly serialized as JSON. Related efforts, such as the contemporary redevelopment of existing text coordinate systems employed by technologies such as TEI, promise consolidation of a single approach to literature annotation targeting using the W3C's Web Annotation Data Model (WADM).

Domain-specific tools for creation and dissemination of new copyright-free scientific data from existing research literature have been refined during the past decade in fields such as taxonomy. Data resources such as the Biodiversity Literature Repository (BLR) now support publicly-available analyses designed for machine consumption of tens of thousands of publications. Biodiversity has been at the forefront of this development because of the circumstances of species loss, which means that assessment of human impact on populations must rely on historical publications. In particular, taxonomy has historically employed concise species description methods which can be readily encoded using formal schemes. Techniques developed by Plazi for analyzing such literature now enable this process to proceed at scale since they are biased towards full automation, with delay associated only with human data quality control. Following legal review, it has been established that such records—we will refer to them as ‘scientific treatments’ derived from existing publications—are not subject to copyright restrictions, which may apply to the original published literature. This development has already transformed biodiversity research methods: hundreds of thousands of treatments and material citations have been reused by the Global Biodiversity Information Facility (GBIF) and the Swiss Institute of Bioinformatics Literature Services (SIBiLS) and included in new publications across the life sciences. Critical cross-domain research e.g. understanding relationships between habitat loss and virus mutation depends on improving such automation, because of the scale and access difficulties with historic literature, as well as the need for rapid response in situations such as the COVID crisis.

However until recently, although they already form FAIR Zenodo records in the case of BLR, these treatments could not be directly connected via annotation to the publications from which they were derived. They were contained only in the IMF files generated by Plazi’s GoldenGate Imagine environment. Consequently, functionality for detailed scrutiny and maintenance of such scientific treatments has been restricted—leading to limitations on reuse and preservation. Dependence on the GoldenGate processing workflow also constrained ongoing further enrichment by the research community. Considering the millions of pages of published literature from which data still has to be liberated in the biodiversity domain alone, standards-based access to scientific treatments is essential.

During March 2021 a collaboration between the National Museum of Natural History in Paris, which is one of the publishers (along with 9 other institutions) of the European Journal of Taxonomy (EJT), Plazi and Data Futures, commenced transformation of existing scientific treatments of EJT articles previously created using GoldenGate to generate WADM annotations. This project addresses a special case, since EJT is a diamond open access journal, though most key gains translate equally to literature which is subject to copyright restrictions. Two key technologies are being employed: the International Image Interoperability Framework (IIIF) and the Annotation Collection data type supported by the Zenodo repository. IIIF developed from requirements in the medical community and in the social sciences and humanities, and its consortium currently comprises 132 institutions, and thousands of IIIF implementations internationally are now operating. As a result multiple large-scale Free and Open Source (FOSS) initiatives have developed IIIF-compatible research tools, including Mirador and Universal Viewer. Moreover, the IIIF presentation API provides comprehensive support for WADM annotations. The Zenodo Annotation Collection data type cements WADM annotations, such as those generated by Data Futures using the components of Plazi's treatments, to literature page data from IIIF services via PIDs. This significantly increases discovery of research that has been output as annotation, and enables a wide range of applications (including spreadsheets and websites, as well as more specialized research tools) to consume scientific treatments automatically. Together, these developments make scientific treatments browseable and maintainable using FOSS applications. The preservation robustness bestowed by trusted repositories such as Zenodo, as well as their discovery functionality, such as metadata harvesting APIs, Elasticsearch and PIDs radically increase the long-term reliability and reuse value of annotation.



Plazi, a394:

Emphysemastix dracarys Olsen & Enghoff sp . nov . urn : lsid : zoobank . org : act : 3 EE 9 B 143 –...

treatment

Plazi, a203:

Emphysemastix dracarys Olsen & Enghoff...

taxonomicName | class: Diplopoda | family: Gomphodesmidae | genus: Emphysemastix | kingdom: Animalia | order: Polydesmida | phylum: Arthropoda | species: dracarys

European Journal of Taxonomy 67

6. Body width 9.0–9.3 mm. Paxill
and L relatively short, straight
– Body width 13.1 mm. Paxillus
M and L longer, more curved (

Emphyse
urn:lsid:zoobank.or

Diagnosis

Differs from the other species of
subtriangular ventral lobe on the
developed process L, and having in

Etymology

The name is to be treated as a noun in apposition. The tip of the gonopod looks like the gaping mouth of a dragon when seen from the dorsal side. The word "dracarys" is a command used in the TV series "Game of Thrones" to make dragons breathe fire.

Material examined

Holotype

TANZANIA • ♂; Iringa Region, Iringa City; 7°46′ S, 35°42′ E; Mar.–Apr. 1996; L. Sørensen leg.; NHMD 621675.

Other material

TANZANIA • 1 ♀; same collecting data as for holotype; NHMD 621676.

European Journal of Taxonomy publication processed with WADM annotations

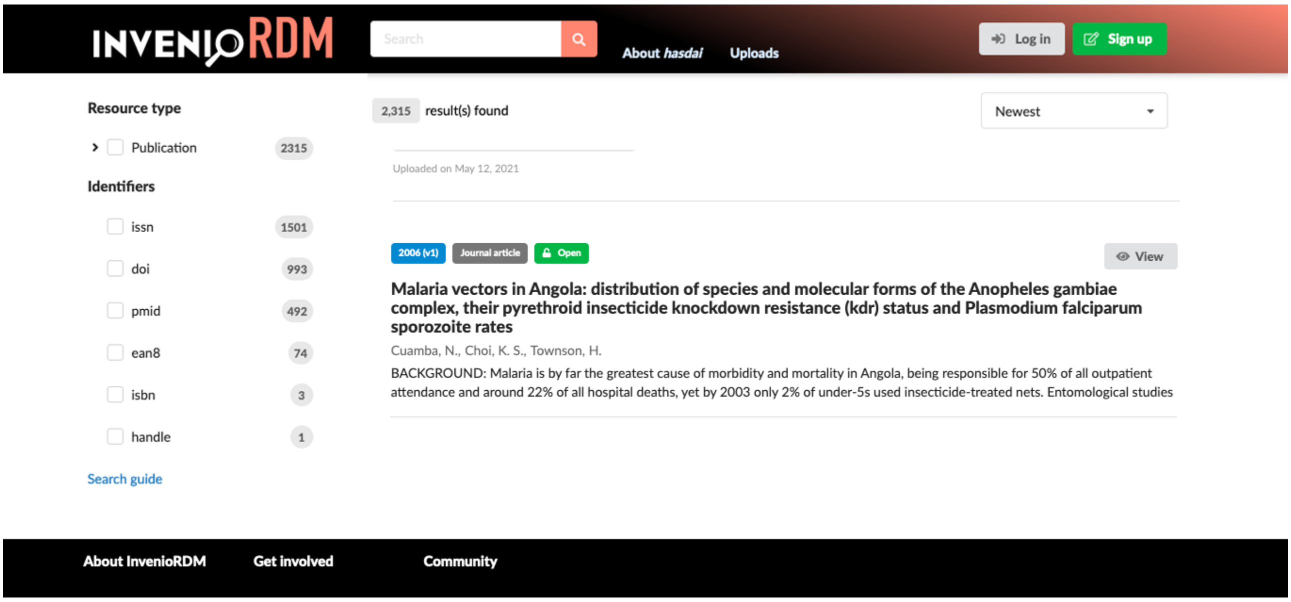
In the illustration on the previous page, taxonomic treatment ("treatment") and taxonomic name ("taxonomicName") annotations derived from a EJT article (https://doi.org/10.5852/ejt.2020.675) which were previously processed into Plazi's IMF format, and already deposited as a Zenodo record http://doi.org/10.5281/zenodo.4332927 are now transformed into WADM annotations against a IIIF service—displayed here using the Mirador IIIF FOSS, and generally accessible for example using the Universal Viewer IIIF FOSS: http://universalviewer.io/uv.html?manifest=https://ejt.biodiversity.hasdai.org/11570/manifest.json#?c=0&m=0&s=0&cv=0&xywh=-1260%2C-125%2C4105%2C2494.

This project has led to the creation of an annotated EJT IIIF data service operated by the *hasdai* partnership with CERN which, since EJT is an open publication, provides a new unrestricted reading interface for browsing the journal. However, the EJT-IIIF also creates new opportunities for reuse and enrichment of scientific treatments produced using the Plazi IMF format generated by GoldenGate. By creating annotations in multiple standards-based representations, including WADM, the existing scientific treatment components can be visualized and edited in their original context for the first time, and new annotations can be created. In addition, augmenting existing BLR/Zenodo records with annotation data enables the IIIF service to deliver individual page fragments of the publications interactively to external applications. Implementation of native IIIF support by InvenioRDM (scheduled in early 2022 by Zenodo) will create the foundation for microservices to deliver IIIF page fragments corresponding to scientific treatment annotations for communities beyond the Biodiversity Literature Repository. Further work on credential management is necessary, but such annotation infrastructure also has the potential to support automatic versioning of scientific treatment Zenodo records—gaining full discovery and preservation benefits for the on-going enrichment of literature.

III. Redelivery of Infectious Disease Literature Repository

VecNet was founded in 2011 as part of the Malaria Eradication Research Agenda (malERA), originally funded by the Bill and Melinda Gates Foundation. Today the number of malaria cases remains between 350 and 500 million people infected worldwide each year, and up to a million cases annually lead to death. The malERA experts concluded that progress with malaria elimination depended on widespread access to, and the means to analyze all the existing research literature relating to malaria. Unfortunately by 2019 VecNet was no longer funded and even it's Datacite repository service ceased. VecNet data saved by the Tropical and Public Health Institute, Switzerland was maintained by Hesburgh Libraries, Notre Dame University as a Fedora repository, but in 2020 this service was also terminated because of funding shortfall for internet security work.

Data Futures imported a Fedora export of VecNet from Notre Dame University in 2019, using the MongoDB-based *freizo* data rescue platform and generated an Invenio3-based VecNet repository (https://vecnet.nd.hasdai.org) as part of the *hasdai* partnership with CERN. Under the InvenioRDM development VecNet has subsequently been made available to the research community and also formed a use-case for transformation at scale of vulnerable literature repositories to improve long-term sustainability. Revisions of this repository, based on recent InvenioRDM releases are preserved at http://vecnet.med.hasdai.org/ (latest at the time of writing: https://may21.vecnet.dev.hasdai.org/). VecNetRDM will form one of the first literature services using InvenioRDM when it is released later in 2021.



Resource type

- □ Publication 2315

Identifiers

□ issn	1501
□ doi	993
□ pmid	492
□ ean8	74
□ isbn	3
□ handle	1

Search guide

2,315 result(s) found

Newest

Uploaded on May 12, 2021

2006 (v1) Journal article Open

View

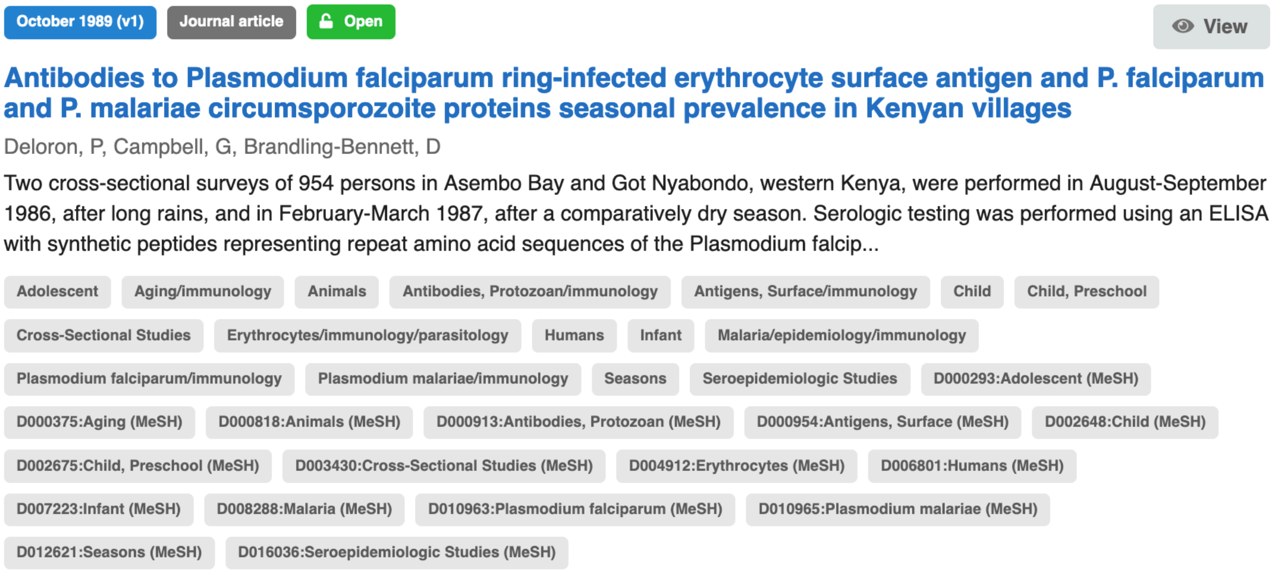
Malaria vectors in Angola: distribution of species and molecular forms of the Anopheles gambiae complex, their pyrethroid insecticide knockdown resistance (kdr) status and Plasmodium falciparum sporozoite rates

Cuamba, N., Choi, K. S., Townson, H.

BACKGROUND: Malaria is by far the greatest cause of morbidity and mortality in Angola, being responsible for 50% of all outpatient attendance and around 22% of all hospital deaths, yet by 2003 only 2% of under-5s used insecticide-treated nets. Entomological studies

VecNetRDM repository generated automatically from historic VecNet data

During January 2021 the Swiss Institute of Bioinformatics provided Data Futures with MEDLINE bibliographic records from its SIBiLS service, which enabled identification of publications occurring in both PubMed and VecNet by comparing 15,567,309 author name occurences. Valuable Medical Subject Heading (MeSH term) metadata could then be extracted, which was either directly provided with MEDLINE records or automatically assigned by SIBiLS. VecNetRDM records were then enriched automatically with these MeSH terms. This malaria-specific literature repository now provides access via PMID as well as DOI, EAN8, ISBN, ISSN and HANDLE using Elasticsearch.



October 1989 (v1) Journal article Open

View

Antibodies to Plasmodium falciparum ring-infected erythrocyte surface antigen and P. falciparum and P. malariae circumsporozoite proteins seasonal prevalence in Kenyan villages

Deloron, P, Campbell, G, Brandling-Bennett, D

Two cross-sectional surveys of 954 persons in Asembo Bay and Got Nyabondo, western Kenya, were performed in August-September 1986, after long rains, and in February-March 1987, after a comparatively dry season. Serologic testing was performed using an ELISA with synthetic peptides representing repeat amino acid sequences of the Plasmodium falcip...

Adolescent | Aging/immunology | Animals | Antibodies, Protozoan/immunology | Antigens, Surface/immunology | Child | Child, Preschool | Cross-Sectional Studies | Erythrocytes/immunology/parasitology | Humans | Infant | Malaria/epidemiology/immunology | Plasmodium falciparum/immunology | Plasmodium malariae/immunology | Seasons | Seroepidemiologic Studies | D000293:Adolescent (MeSH) | D000375:Aging (MeSH) | D000818:Animals (MeSH) | D000913:Antibodies, Protozoan (MeSH) | D000954:Antigens, Surface (MeSH) | D002648:Child (MeSH) | D002675:Child, Preschool (MeSH) | D003430:Cross-Sectional Studies (MeSH) | D004912:Erythrocytes (MeSH) | D006801:Humans (MeSH) | D007223:Infant (MeSH) | D008288:Malaria (MeSH) | D010963:Plasmodium falciparum (MeSH) | D010965:Plasmodium malariae (MeSH) | D012621:Seasons (MeSH) | D016036:Seroepidemiologic Studies (MeSH)

MeSH terms added automatically using the SiBILS Service

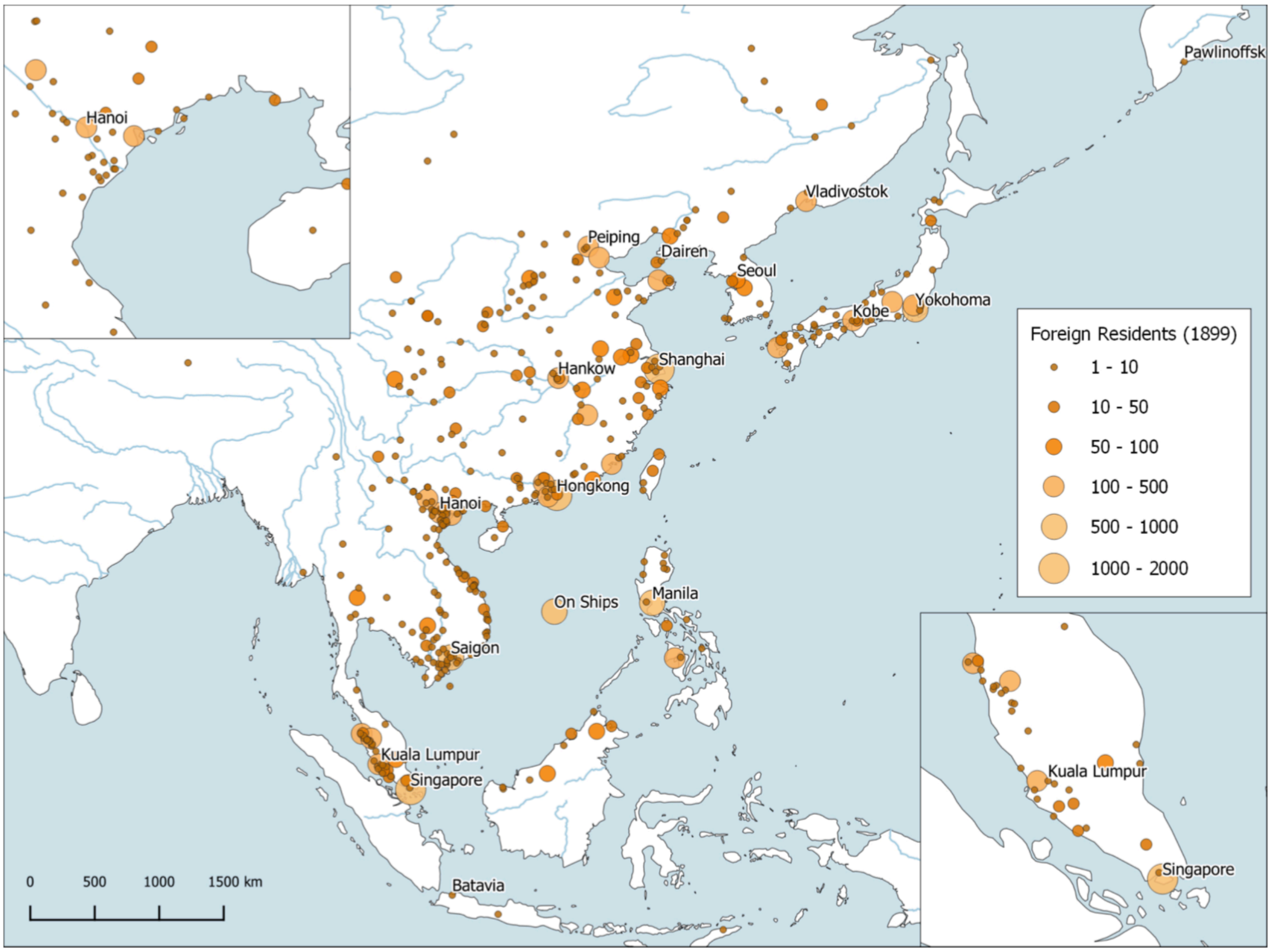
IV. Preserving Data Resources in the Social Sciences and Humanities

Redelivery and enrichment of existing research data resources has little value if, in their turn, such efforts become vulnerable to the same technology obsolescence and institutional change within a decade. Regrettably, even shorter lifetimes have become accepted for data not actually published in journals, and this precludes almost all analysis preservation. Scientific publishers' business models lock investment which has usually been created in publicly-funded activities behind paywalls with repressed discovery services. The Findable, Accessible, Interoperable and Reusable (FAIR) Principles seek to improve this situation, though compliance has been slow and inconsistent. Emerging techniques for automating creation of scientific treatments using machine learning methods promise to extend the creation of discoverable, copyright-free scientific facts from literature at scale in domains other than the life sciences. However, effective preservation of these outputs will be critical.

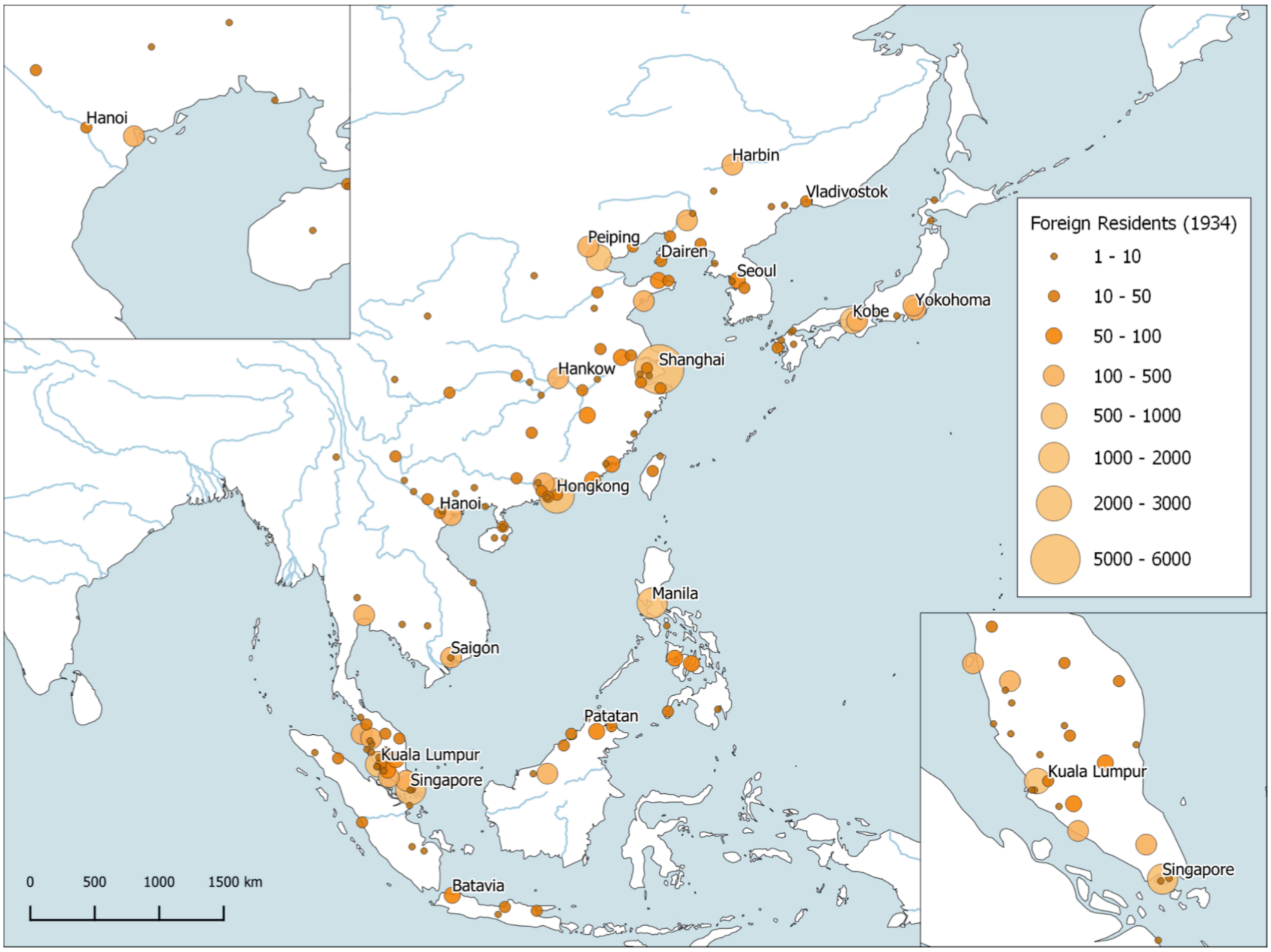
This case study addresses creation of datasets designed to be reused in multiple ways by the research community, and it evaluates preservation and discovery mechanisms for research output as annotation. Collaboration between the Institute for European Global Studies, Basel, and Data Futures processed listings of foreign residents in the 1896, 1899 and 1934, 1937 volumes of the Asian Directories & Chronicles serial, which was published annually by The Hong Kong Daily Press between 1863 and 1941. The years featured in this dataset were selected because of their relationship to historic events in East Asia. The First Sino-Japanese War, waged from July 1894 to April 1895, was followed by establishment of large numbers of small communities of foreign residents throughout East Asia. In the early 20th century consolidation of foreign residents in larger communities in coastal cities was followed by a marked exodus during escalating conflict in the Second Sino-Japanese War between July 1937 and September 1945, and its origins in the Japanese invasion of Manchuria in 1931. European resident population shifts based on the years covered by this dataset are striking when rendered geographically. Note that while very small groups of foreign nationals are still registered in data collected 35 years later, the scale measuring coastal populations registers far more individuals in the visualizations below.

With the current exceptions of 1866, 1867, 1872, 1875 and 1884 all of the volumes of the Directories & Chronicles have been assembled in a single *freizo* digital corpus from which an Invenio repository has been generated. Digitization of the pages of the volumes, creation of a IIIF service and analysis of OCR data has enabled automated detection of each person record in the foreign resident listings contained in the serial, and generation of 60,712 such annotations from this 4-year sample. The OCR text has subsequently been corrected with the aid of surname and location dictionaries created from the corpus, and searchable person datasets (individuals’ name occurrences) have been generated, supported by a JSON schema. This dataset is self-describing to promote long-term technology-independent accessibility. Together these components form a Zenodo record at https://doi.org/10.5281/zenodo.2580997. Inclusion of the schema means that the foreign resident person occurence data can be consumed efficiently by a range of existing and potentially future research tools. This record includes sample geographic visualizations rendered by QGIS, which show the location indicated in the serial for each person during the four years in question, two of which are reproduced here.

Employing Zenodo's Annotation Collection dataset type significantly improves automated discovery. The current version of the record employs OADM—a pre-cursor of WADM which is supported by existing IIIF viewers such as Mirador. When processing of the remaining years of the serial is complete, more than 800,000 person occurrences will be made accessible in the same way, providing access to each occurrence as JSON, organized according to the name/ location schema plus JSON annotations, and linked to original locations on pages of the serial via IIIF. Both annotations and person occurrence data are also available as an Invenio corpus repository (having an optimized data model), providing Elasticsearch for person, year and location terms.



Zenodo dataset providing names and locations of individual foreign residents registered in East Asia during 1899, rendered using the QGIS FOSS



Zenodo dataset providing names and locations of individual foreign residents registered in East Asia during 1934, rendered using the QGIS FOSS

IV. Summary

Advances in machine learning are rapidly improving the volume and accuracy of facts that can be extracted from digitized sources. Publishing of such metadata and its preservation in the long-term—including its authority and potential for future enrichment based on persistent links to sources—is crucial for the global research community. Web Annotation Data Model and the International Image Interoperability Framework are key technologies which promote reuse of analyses of scientific literature. Moreover, annotation collections can form training datasets for new domain-specific analysis tools such as neural networks. WADM enables convergence of existing investments which have already extracted copyright-free scientific treatments from the vast resources of historic publications. However, these building blocks do not themselves guarantee long-term accessibility; rather, trusted repositories are essential for providing stable infrastructure for and comprehensive discovery of research data.

Redelivery of existing treatments of taxonomic literature as standards-based annotations is an important step towards establishing machine- as well as human-accessible access to the estimated 500 million pages of biodiversity literature and its reuse in multiple domains.

New repository technologies which are trusted in the sense of financial sustainability as well as technical utility, provide long-term infrastructure for vulnerable historic scientific literature investments and enable them to be reused and further enriched. But identifying and regaining access to such resources before they are finally lost remains a daunting challenge.