RDA IG Preservation Techniques, Tools and Policies

Preserving Scientific Annotation Working Group (PSA-WG)

Case Statement Revision

---

1. Background

Annotation has become an established research tool during the last 5 years. However, since 2016 rapidly-increasing production of annotations using a few technologies has developed into a much more complex situation. On one hand machine learning is now revolutionizing the creation of annotations for a much wider range of purposes. Also, existing analyses of literature—producing data not previously identified as annotations—are now being converted into standards-based annotations: new primary research objects having independent discovery methods (for example see European Journal of Taxonomy case study, Annex-A). On the other hand, mainstream commercial software applications are adding annotation functionality to their products as integrated services, producing local and sometimes proprietary rather than independently consumable annotation data. This situation poses new questions for policy-making, affecting preservation and reuse. For example, whether annotation workflows permit *in practice* the production of research data which is reusable outside the immediate project environment is seldom self-evident. Re-thinking the activities of an RDA Preserving Scientific Annotation Working Group is therefore required.

---

2. Activities

Challenges arising from the underlying vulnerability of stand-off annotation and the need to communicate the need for practices promoting preservation and reuse of annotation, as originally identified by PSA-WG at its BoF in 2018, still remain. We do not describe these underlying issues in detail here, but attach our earlier document as Annex-D. The 2021-2123 work-plan for PSA-WG proposed here focusses on establishing an RDA information service about scientific annotation infrastructures: in addition to communicating successful preservation & reuse of annotations via case studies, PSA-WG will actively gather and document the spectrum of emerging applications which generate or employ annotations as reports for RDA members. This will address the vacuum of comparative information about annotation infrastructures which persists both across scientific domains and internationally. PSA-WG will document emerging annotation technologies and vulnerabilities and, in doing so, potentially enable RDA to develop a leadership role in this field.

Preservation Tools Technologies & Policies (PTTP-IG) has built a foundation for this work during the last 3 years. In 2020 it delivered webinars through collaboration with the Frankfurt Book Fair organization and with RDA's Research Data Management in Engineering IG (RDMinEng-IG) to highlight annotation issues in the publishing and engineering sectors respectively. PTTP has also worked together with existing annotation projects in the biodiversity, medical and social science and humanities communities and these projects are described in Annexes A through C. PSA-WG will communicate these new activities to the RDA community via the RDA website and it will deliver three further webinars using this material—one announced at and delivered shortly after VP18 and the second during Spring 2022.

3. Preservation and Communication

PSA-WG will create and maintain a dedicated InvenioRDM repository, operated and funded by the *hasdai* partnership with CERN in order to disseminate and preserve reports and webinar material arising from its activities (preliminary materials form Annexes A-C hereto, and potentially further projects will be added later in the lifetime of the WG). Where possible such materials will additionally be installed on or linked with the RDA website. The PSA-WG repository will also support hands-on demonstration of tools and technologies for preserving and reusing research investment producing annotation. The PSA-WG repository will provide the primary vehicle for communication about the spectrum of emerging applications which generate or employ annotations, and potentially social media posts could also be generated from repository materials.

Use of an independent repository supports both RDA's policy of 'member access only' and advanced functionality for demonstrating effective long-term preservation of annotation data. InvenioRDM has been selected because it supports annotation data and it will form the technology base for forthcoming enhancement of Zenodo, OpenAIRE's repository at CERN, which supports research data efficient discovery by the wider community. InvenioRDM's use of Elasticsearch means that searching within the PSA-WG repository would be comprehensive without demanding rigorous tagging. The preliminary materials presented in the Annexes hereto show that associated records have also been generated in Zenodo to achieve wider discovery. The PSA-WG repository would employ authentication via ORCID plus a white-list generated from the RDA membership.

4. Outputs

In summary, PSA-WG will:

- create and operate the PSA-WG repository and maintain RDA membership access via ORCID-based white-list authentication;
- deliver a series of webinars based on case studies, highlighting successful annotation infrastructures—materials from which will made available via the RDA website and preserved in the PSA-WG repository;
- produce reports surveying emerging applications which generate or employ annotations and the WG will communicate them to the membership via the RDA website and the PSA-WG repository;
- publish recommendations for the RDA membership, drawn from PSA-WG case studies and reports after consultation with TAB

Existing collaborations with RDA Interest Groups and Working Groups will be further developed and collaborations with other Groups—potentially leading to another tier of case studies—will be sought. In particular, PSA-WG will extend credentials for creation and enrichment of materials available in its repository to collaborating RDA Groups.

5. Roadmap

If approved by TAB, PSA-WG would commence building its Invenio repository during Summer 2021 and create an ORCID whitelist permitting reliable authentication by the existing RDA membership before VP18. Since a membership snapshot provided by Secretariat will have to be employed for the whitelist, and new members are expected to need access between this snapshot and the plenary, on-boarding for new members together with approval delegation will also be organized. Initial repository contents would be limited to materials associated with the three new case studies previewed in Annexes A-C hereto and reports produced from studies of emerging annotation infrastructures. During 2022 access for creating and maintaining the PSA-WG repository would be extended to other RDA Groups.

A PSA-WG session application for VP18 in virtual format would be submitted and the first PSA-WG webinar announcement will be made over 3-18 November 2021, with likely delivery date in late November or early December 2021. This webinar would address collaboration with the National Museum for Natural History, Paris, in respect of the new WADM/IIIF data resource for the European Journal of Taxonomy. Based on that advertising and delivery experience and dates of subsequent mainstream RDA events, the second webinar would be scheduled during Spring 2022 and the third in early Fall 2022. Depending on the success of these webinars and progress establishing new case study projects and collaborations with other RDA Groups, the program for 2022/3 would be finalized before the end of 2022.

In parallel with its webinar program PSA-WG would conduct a series of investigations into new and emerging annotation infrastructures, focussing in particular on annotations generated using machine intelligence-related applications. The first report in this series would appear before the end of 2021 and is expected to focus on research applications of technology developed by Clarifai, Inc.

Together, these activities would raise awareness within RDA of the potential for discovery and reuse of research investment via annotation infrastructure. Commencing with iPRES in October 2021, PSA-WG will also communicate these outputs beyond RDA, establishing the Alliance as a recognized authority in this field.

Annex-A

Enriching copyright-free bio-taxonomy data derived from existing literature.

Over the last 20 years 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. This process, which historically was not conspicuously called 'annotation', augments bulk machine-based automated discovery of facts reported in publications.

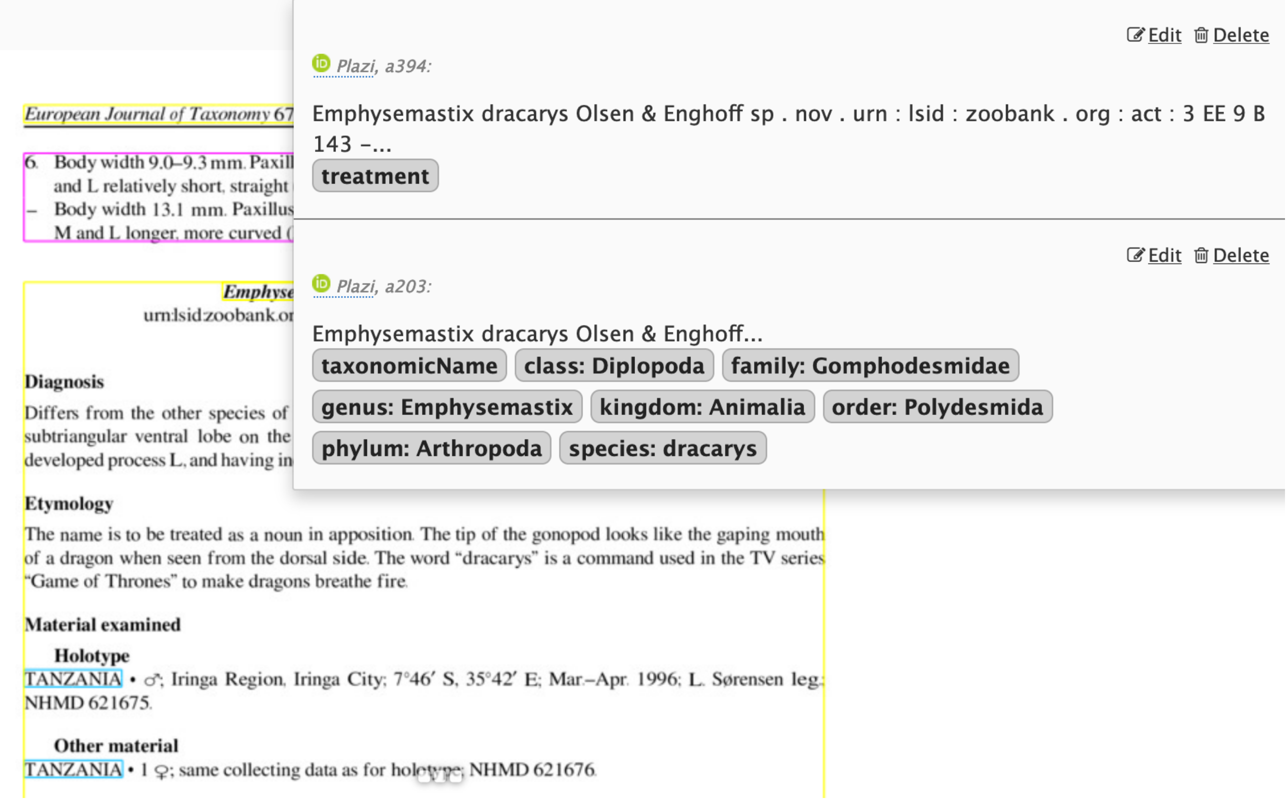
Identifying sections of literature for this purpose is referred to as target definition: the contemporary redevelopment of existing text coordinate systems employed by technologies such as TEI, promises consolidation of a single approach to literature annotation targeting using the W3C's Web Annotation Data Model (WADM).

Creation and dissemination of new copyright-free scientific data from existing research literature has 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.

In the illustration on the next 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://



Edit Delete

Plazi, a394:

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

treatment

Edit Delete

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.

ejt.biodiversity.hasdai.org/11570/manifest.json#?c=0&m=0&s=0&cv=0&xywh=-1260%2C-125%2C4105%2C2494.

European Journal of Taxonomy publication processed with WADM annotations

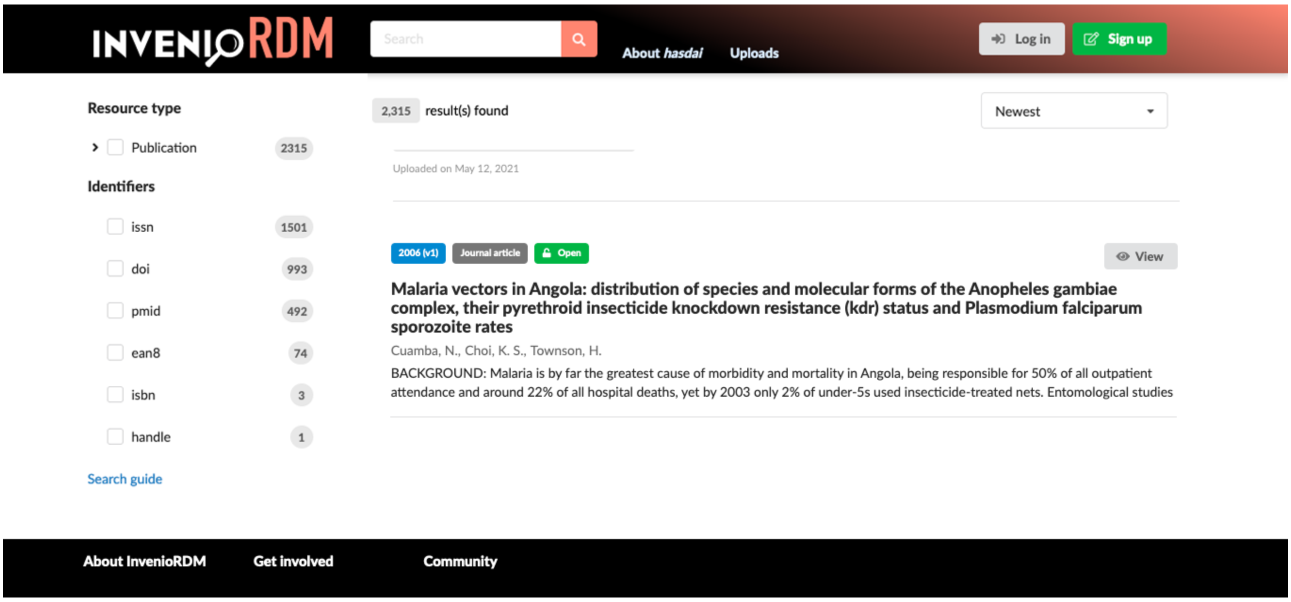
This project has led to the creation of an annotated EJT IIIF data service operated by the *hasdai* partnership with CERN which, since EJT is an open publication, provides a new unrestricted reading interface for browsing the journal. However, the EJT-IIIF also creates new opportunities for reuse and 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.

Annex-B

Redelivery and Enrichment of Infectious Disease Literature Repository

Based on an existing literature repository, this case study demonstrates the potential for creation of annotations via analysis of existing metadata. 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.

A Fedora export of VecNet from Notre Dame University in 2019, was processed using the MongoDB-based *freizo* data rescue platform and an Invenio3-based VecNet repository (https://vecnet.nd.hasdai.org) generated 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 have been preserved, but at the time of writing: https://may21.vecnet.dev.hasdai.org/ is current. VecNetRDM will form one of the first literature services using InvenioRDM when the latter 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

Uploaded on May 12, 2021

2006 (v1) Journal article Open

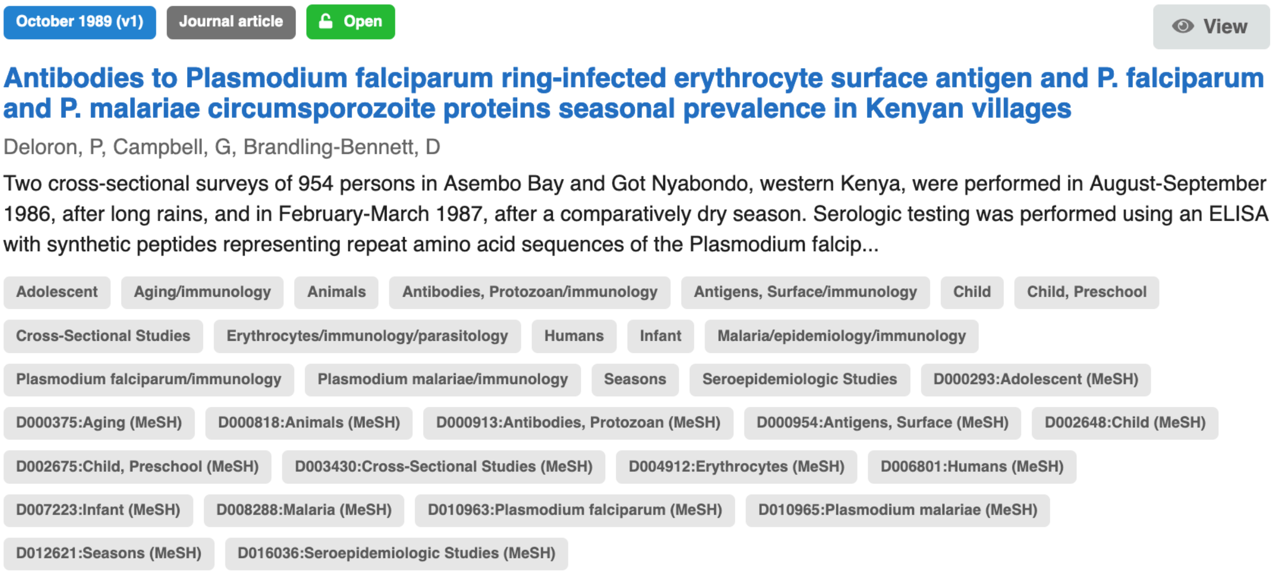
Malaria vectors in Angola: distribution of species and molecular forms of the Anopheles gambiae complex, their pyrethroid insecticide knockdown resistance (kdr) status and Plasmodium falciparum sporozoite rates

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

The screen shot on the previous page shows a record in the VecNetRDM repository generated automatically from historic VecNet data. VecNetRDM can now be searched using MeSH (PMID) identifiers, as well as DOI, EAN8, ISBN, ISSN and HANDLE using Elasticsearch.

During January 2021 collaboration with the Swiss Institute of Bioinformatics provided MEDLINE bibliographic records from the 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 have now been extracted, which were either directly provided with MEDLINE records or automatically assigned by SIBiLS. VecNetRDM records were then enriched automatically with these MeSH terms. Since the MeSH terms are related directly to occurrences in the literature it is now feasible to detect these entities automatically and create WADM annotations against a VecNet IIIF service. The significant enrichment of the literature achieved in this way is indicated by the proportion of MeSH terms now available for the augmented records:



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 using the SiBILS Service

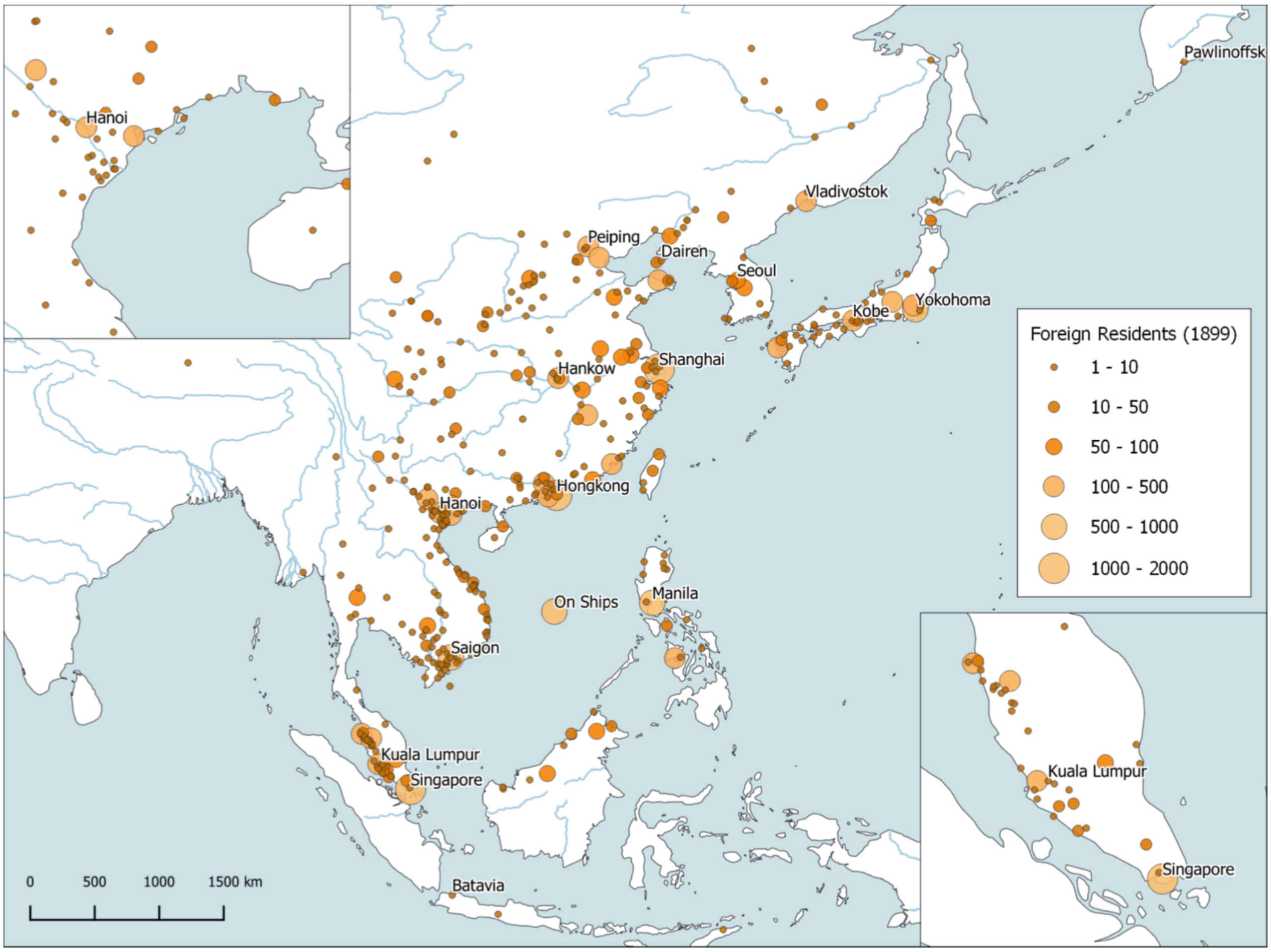
Annex-C

Creation of FAIR Annotation-based Data Resources in the 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 Directories, 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 instance data can be consumed efficiently by a range of existing and potentially future research tools—for example geographic visualizations showing the location indicated for each person during the four years in question are included in the Zenodo record. Attachment of the annotation data using the Zenodo Annotation Collection datatype means that this dataset can be consumed efficiently by external interactive applications. The current version of this record employs OADM—a pre-cursor of WADM which is supported by current IIIF viewers such as Mirador. Both annotations and person occurrence data are also available as an Invenio corpus repository, providing Elasticsearch for person and location terms, and IIIF location of person occurrences on pages of the serial where they are listed.



Zenodo annotation dataset providing names and locations of individual foreign residents registered in East Asia during 1899, rendered using QG