Biodiversity
Heritage
Library

Spring 2023

DRAFT

Unifying Biodiversity Knowledge to Support Life on a Sustainable Planet

A Wikimedia Projects Whitepaper by JJ Dearborn

Major support and hosting for BHL is provided by Smithsonian
Libraries and Archives

To my mind, there is perhaps no better demonstration of the folly of human conceits than this distant image of our tiny world. To me, it underscores our responsibility to deal more kindly and compassionately with one another and to preserve and cherish that pale blue dot, the only home we've ever known."

— Carl Sagan

*Cover image: **Earthrise over Compton Crater**, NASA, Goddard Space Flight Center, Arizona State University*

About BHL

The Biodiversity Heritage Library (BHL) is the world’s largest open access digital library for biodiversity literature and archival materials. BHL is revolutionizing global research by providing free, worldwide access to knowledge about life on Earth. BHL's digital collection provides free access to over 60 million pages from the 15th-21st centuries.

Since its launch in 2006, BHL has served millions of people in over 240 countries and territories around the world. Through ongoing collaboration, innovation, and an unwavering commitment to open access, the Biodiversity Heritage Library continues to transform research on a global scale and provide researchers with the information and tools they need to study, explore, and conserve life on Earth.

BHL operates as a worldwide consortium of natural history, botanical, research, and national libraries. Major support and hosting for BHL is provided by the Smithsonian Libraries and Archives (SLA). To learn more about us, visit https://about.biodiversitylibrary.org

Author

JJ Dearborn, BHL Data Manager | dearbornjj@si.edu

Contributors

The following groups of people provided feedback, insight, and wisdom that made this white paper possible. Thank you all for your time, energy, and thoughtful input. Contributors are listed alphabetically by last name.

Interviews

Grace Costantino (Independent), Diana Duncan (BHL/Chicago Field Museum), James Hare (Internet Archive), Siobhan Leachman (Independent Wikimedian), Andy Mabbett (Independent Wikimedian), Katie Mika (Harvard University), Dr. Rod Page (University of Glasgow), Carolyn Sheffield (The Nature Conservancy), Alex Stinson (Wikimedia Foundation), and Andra Waagmeester (Micelio/Independent Wikimedian).

Peer Review

Bianca Crowley (SLA/BHL), Jacqueline Chapman (SLA), Sandra Fauconnier (Independent Wikimedian), Giovanna Fontenelle (Culture and Heritage, Wikimedia Foundation), Colleen Funkhouser (SLA/BHL), Nicole Kearney (Museums Victoria/BHL Australia), Martin Kalfatovic (SLA/BHL), Siobhan Leachman (Independent Wikimedian), Mike Lichtenberg (BHL), Susan Lynch (BHL/NYBG), Bess Missell (SLA/BHL), Richard Naples, (SLA/BHL), Jake Orlowitz (WikiBlueprint), Dr. Rod Page (University of Glasgow), Suzanne Pilsk (SLA/BHL), Joel Richard (SLA/BHL), Connie Rinaldo (BHL), Fiona Romeo (Culture and Heritage, Wikimedia Foundation), and Jackie Shieh (SLA).

Table of Contents

Executive Summary

Urgent action is needed to combat the impacts of climate change, biodiversity loss, and secure a sustainable future for the planet. Policymakers, national governments, and intergovernmental organizations rely heavily on data-driven research to inform the key environmental indicators and policies required to meet this moment of crisis. The need for the global biodiversity community and its disparate data silos to build a unified biodiversity knowledge graph rich in human and machine-curated interlinkages has never been greater.

As a prominent member of the global biodiversity informatics community, the Biodiversity Heritage Library (BHL) has a central role to play. BHL data:

1. improves scientific understanding of ecological change, deeper into time, at both global and hyper-local scales;
2. assists decision-makers in shaping global environmental policy informed by the historical record; and
3. bridges knowledge gaps and facilitates information exchange regarding our planet's history.

By liberating over 500 years of data and making it open and Findable, Accessible, Interoperable, and Reusable (GO FAIR initiative, 2020), BHL fosters universal bioliteracy and meets global climate challenges. In response to urgent calls from the scientific community to de-silo and connect disparate climate- and biodiversity-related datasets, BHL is investigating Wikimedia's core projects, in particular, Wikidata and its underlying architecture Wikibase. Wikidata has emerged as a global information broker and collaboratively edited data store that provides a unified technical infrastructure well-positioned to support biodiversity and climate science, environmental policy, and global efforts to monitor the health of our planet.

BHL's evolution from a digital library to a big data repository requires normalizing, standardizing, and enriching data at a scale that equates to lifetimes of manual labor. To meaningfully contribute to the vital work of global biodiversity data infrastructure, BHL must embrace computational approaches that will hasten the semantic enrichment and rapid dissemination of data in its corpus. BHL must pursue new capacity-building partnerships and ensure its resources match its mission, in order to support the scientists and policymakers working towards life on a sustainable planet.

Background

Headquartered at the Smithsonian Libraries and Archives (SLA) in Washington, D.C, the BHL Secretariat commissioned the writing of this Wikimedia white paper for BHL's consortium partners, its core users, and members of the open knowledge movement. Research into the Wikimedia information ecosystem began in September 2021.

This white paper provides concrete use cases and recommendations to help BHL navigate Wikimedia project investments strategically. (See Appendix 1: Summary of Recommendations) Given the BHL Consortium's current capacities, strategic investment decisions will need to be made. With additional resources, BHL could expand its efforts to meet pressing global challenges.

The introductory material provides a thorough examination of BHL's evolution from a content provider to a biodiversity data knowledge base so that BHL's primary audiences may more fully understand:

- the unique contributions BHL has to make to the global biodiversity data infrastructure through the lens of climate change, species loss, and knowledge representation,
- evolving linked open data standards, principles, and protocols, and
- BHL's three big data challenges:

  1. Correcting and Transcribing OCR Text

  2. Improving BHL Search Precision and Retrieval

  3. Linking and Depositing BHL Data with Global Knowledge Bases.

Because BHL is a digital library focused on improving its data management strategies, special focus has been paid to Wikimedia's core projects and global volunteer community whose mission is to "empower and engage people around the world to collect and develop educational content under a free license or in the public domain, and to disseminate it effectively and globally" (Wikimedia Foundation, 2018).

Aligned in spirit, the Biodiversity Heritage Library embraces this mission.

Purpose

After a careful review of BHL's past involvement in Wikimedia projects (Costantino, 2015), the BHL Secretariat recognizes that there are more opportunities for improving BHL's data flow downstream through informal and formal GLAM-Wiki partnerships ("galleries, libraries, archives, and museums" with Wikimedia).

A subset of Wikimedia projects and tools were explored for their potential to further BHL's current strategic plan and maximize contributions to the global biodiversity infrastructure. Use cases were identified through a review of the scholarly literature, stakeholder interviews, and virtual attendance at relevant workshops, committee meetings, and conferences. Below is the list of use cases formally identified for BHL's further involvement in Wikimedia projects.

Identified Use Cases

Wikidata

1. Name Disambiguation
2. Adding BHL Articles and DOIs to Wikidata
3. Enabling Interdisciplinary Research with SPARQL

Wikibase

4. Extending BHL's Data Model
5. Wikibase Front-end Showcase

Wikimedia Commons

6. Image Search with Structured Data Commons (SDC)
7. Image Use Metrics and Analytics

Wikisource

8. Handwritten Text Transcription

Wikipedia

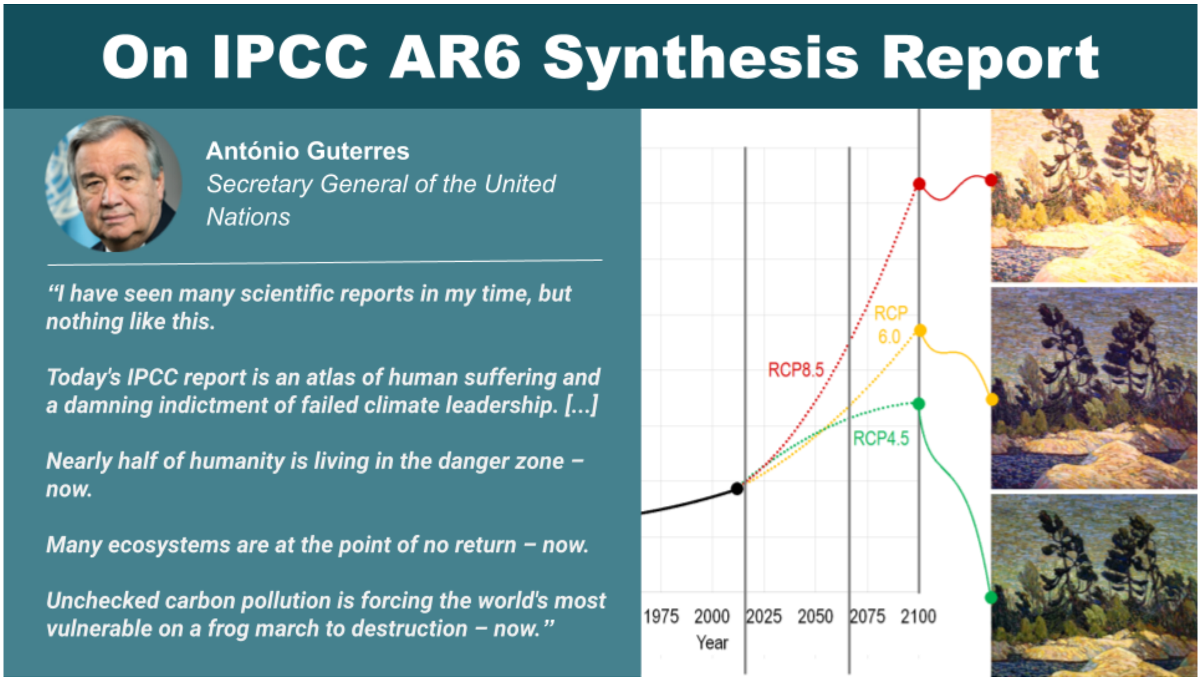
9. Increasing User Traffic to BHL
10. Bridging Knowledge Gaps with Campaigns

This list of use cases found is only the starting point. The hope is that members of this community are inspired to experiment, share, and find new use cases in the Wikimedia information ecosystem.

1.0 Introduction

1.1 Biodiversity Data for our Planet

In April 2021, the Intergovernmental Panel on Climate Change (IPCC) released its *Sixth Assessment Report* which to date, represents the most comprehensive summary of the physical science of climate change. The 3,949-page report collates over 14,000 research studies and shows that human activity is responsible for global mean temperature change. (IPCC Report: 'Code Red' for Human Driven Global Heating, Warns UN Chief, 2021) The assessment's alarming conclusions are a result of "improved data on historical warming" (United Nations, 2021). The recent *IPCC AR6 Synthesis* report reiterates that the time for action is now.



On IPCC AR6 Synthesis Report

António Guterres
Secretary General of the United Nations

"I have seen many scientific reports in my time, but nothing like this.

Today's IPCC report is an atlas of human suffering and a damning indictment of failed climate leadership. [...]

Nearly half of humanity is living in the danger zone – now.

Many ecosystems are at the point of no return – now.

Unchecked carbon pollution is forcing the world's most vulnerable on a frog march to destruction – now."

Figure: UN Secretary Guterres states that "now" is the moment to act and a "quantum leap" is required. The most recent IPCC AR6 Synthesis Report represents a final warning to humanity. (United Nations, 2023)

Gregory Nemet, IPCC Author and Professor at the University of Wisconsin-Madison emphasizes the need to ramp up carbon removal efforts from current rates of two gigatons to eight gigatons per year within the next ten to fifteen years and states that "the goal of 1.5 °C degree warming is at great risk – we are 1.2 °C degrees; 1.5 °C is coming fast and furious by 2030." (Nemet et al., 2023) The collective failure to decarbonize our atmosphere will increase the frequency of extreme weather events. Already, the impacts of the triple planetary crisis of pollution, climate change, and biodiversity loss are proving devastating to human health, economic output, food security, coastal and island communities, and terrestrial and marine habitats. (Diffenbaugh & Barnes, 2023) These impacts are not distributed equally; rather, marginalized, disadvantaged, and impoverished communities, both human and organism, remain among the most vulnerable to climate shocks. (Edmonds, 2022)(World's Most Vulnerable Nations Suffer Disproportionately, n.d.)

Despite the bleak outlook, there is still time to change course. Largely off the table, are past predictions that we are tracking against RCP8.5, the high-emissions scenario in which energy consumption relies primarily on coal and its expansion. (Harrisson, 2021) Instead, the focus has shifted from doomsday scenarios to collective action — now — towards decreasing consumption, scaling renewable technologies, and removing excess carbon dioxide (CO_2) from the atmosphere. Forests, mycorrhizal-rich soil, and microalgae are the lungs of our planet acting as carbon sequestration power-houses. Revegetating our planet, bolstering the earth's mature ecosystems, and conserving Earth's biodiversity could mitigate the worst effects of climate change, yet to come. Conventional carbon dioxide removal (CDR) methods dubbed "nature-based," solutions account for almost all (99.9% or 2 $GtCO_2$ per year) carbon removal today. (The State of Carbon Dioxide Removal Report, 2023)

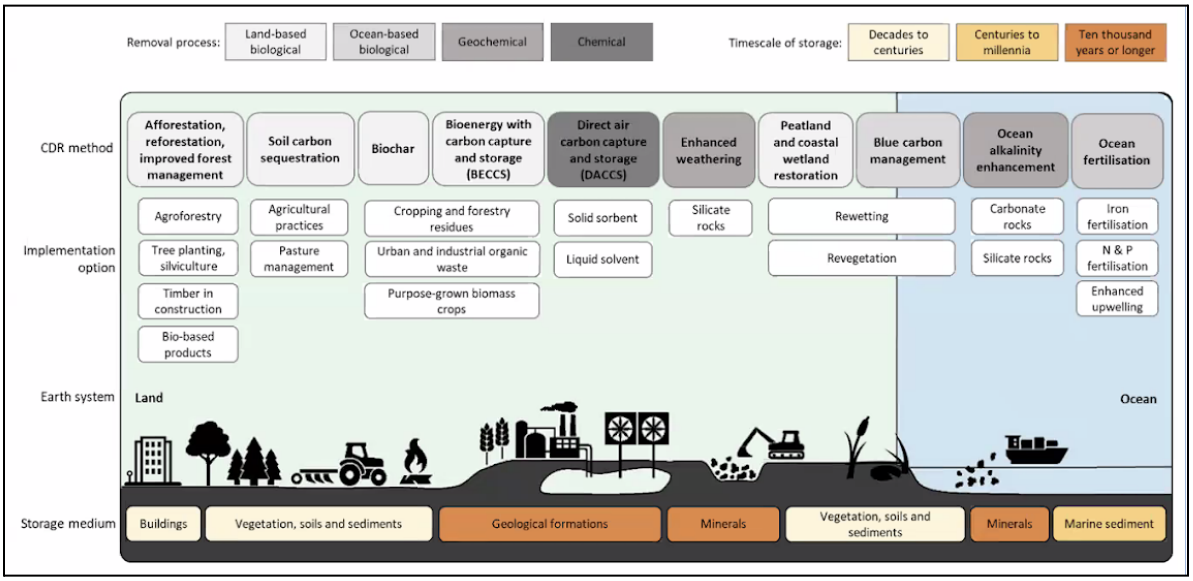


Figure: Novel approaches to carbon removal are emerging and will need to ramp up dramatically to meet Paris Agreement targets. All CDR *methods will need to be deployed strategically and in tandem. (Image: Nemet et al., 2023)*

Alongside carbon removal approaches, standard measurement, reporting, and verification is needed to monitor global progress and incentivize good actors. Data interoperability across datasets, metrics, and systems continues to impede collective action around carbon sequestration goals. Major investments to improve technical infrastructure is required for international coordination. (Nemet et al., 2023) Monitoring the health of the planet is contingent on large inputs of interoperable, clean, openly accessible data from a variety of sources with a shared technical infrastructure to support it.

Central to global biodiversity data aggregation and dissemination efforts is the Global Biodiversity Information Facility (GBIF), which provides free and open access to biodiversity data. Climate change research and subsequent policy recommendations rely on the quality and quantity of data in GBIF. In short, GBIF-mediated data is used to determine environmental policy decisions made by nation-states and intergovernmental organizations (GBIF, 2019). Moreover, two events represented a watershed moment for climate policy:

- The Convention on Biological Diversity (CBD COP-15), and
- The United Nations Climate Change Conference (COP-26).

The development of the post-2020 global biodiversity framework at CBD COP-15 and the adoption and adherence to national emission targets by countries at COP-26 will largely determine the fate of our planet. High-level decision-makers now rely heavily on data-driven research to inform policy:

> *"To effectively conserve biodiversity, it is essential to make indicators and knowledge openly available to decision-makers in ways that they can effectively use them. The development and deployment of tools and techniques to generate these indicators require having access to trustworthy data from biological collections, field surveys and automated sensors, molecular data, and historic academic literature" (Gadelha et al., 2020).*

The Biodiversity Heritage Library has done an exceptional job of serving the needs of researchers. The next challenge for the global consortium will be to rapidly disseminate the 60 million-page textual corpus as structured data on the web. Extract, transform, and load (ETL) pipelines will need to be piloted with a focus on big data brokers: **GBIF and Wikidata.** The further release of BHL's data will:

1. Improve scientific understanding of ecological change, deeper into time, at both global and hyper-local scales;
2. assist decision-makers in shaping global environmental policy informed by the historical record; and
3. bridge knowledge gaps and facilitate information exchange regarding our planet's history.

> *"Having access to historical literature is essential to characterizing what ecosystems used to look like, what species were present, and what peoples' opinions of the health of the ecosystem were like throughout time. Taxonomic literature allows me to see the whole history of a species laid out before me. I rely on [it] to get a glimpse of how these wonderfully diverse ecosystems used to look… before widespread development…[and use] those baselines to evaluate how current conservation measures are succeeding." — Dr. Joshua Drew, Marine Conservation Biologist at The Field Museum of Natural History, Chicago. (Costantino, 2018)*

1.2 An Evolving Data Landscape

In the past decade, open data initiatives have cropped up worldwide under increasing pressure from citizens and governments to make public data freely available on the web. (Attard et al., 2015) In a recent report, "Digitization of the World", it is predicted by the International Data Corporation that the "global datasphere will grow from 33 zettabytes (ZB) in 2018 to 175 ZB by 2025." For reference: one zettabyte is a trillion gigabytes (International Data Corporation, 2018). Information standards, schemas, and open formats are the conduits that have been put in place to help make this deluge of data interoperable, usable, and connected.

In parallel with the open data movement, the field of biodiversity informatics has burgeoned. (Peterson et al., 2015) Computational analysis methods are connecting disparate biodiversity data and allow scientists to infer new patterns about our natural world. A rapidly changing digital landscape means that traditional modes of information delivery and exchange will no longer suffice. New models are now replacing outdated ones.

Despite valiant efforts, eliminating the structural barriers that hinder the true potential of data has proven difficult. Much of the world's data remains locked up, siloed, and underutilized. Several studies estimate that only 1-5% (Burn-Murdoch, 2017) of data is analyzed, and in 2016, IBM reported that over 80% of data is dark data, with that number expected to grow to 93% by 2020 (Trice, 2016).

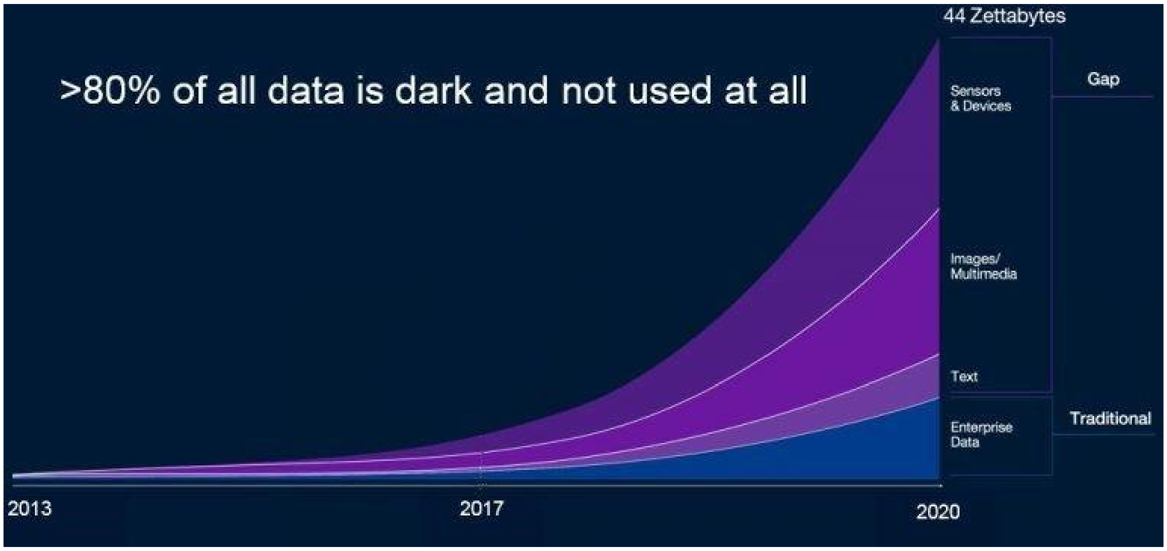


Figure: Unstructured data — "dark data" — accounts for a majority of all data generated. How much untapped potential and hidden knowledge lies within BHL's 60-million-page textual corpus, waiting to be unlocked? (Image: Schroeer, 2017)

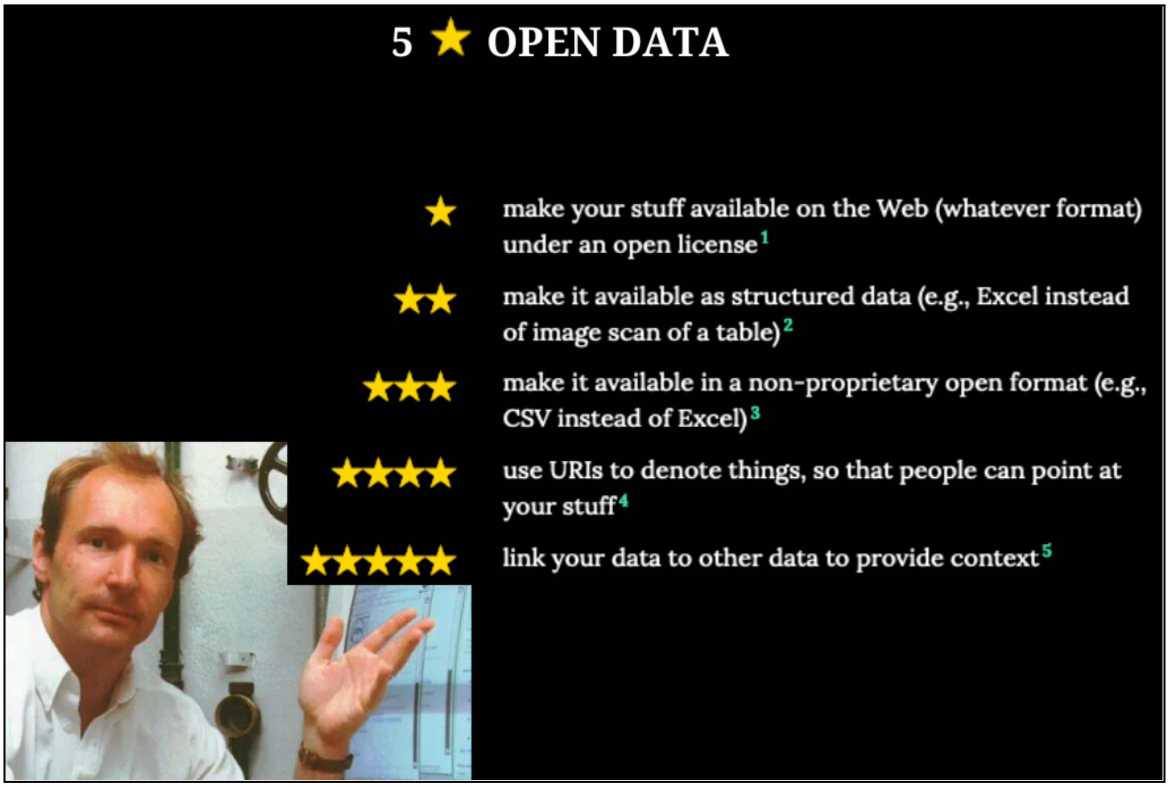
A New Vision For a Smarter Web

In 2001, Scientific American published a pithy article entitled "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." In it, Sir Tim Berners-Lee offers a glimpse into the future where the banal bureaucracy of our daily lives almost disappears, taken care of by computational agents acting on information from the semantic web. In this new information paradigm, the old web has been retrofitted to gain logic and all data is interoperable and connected. Berners-Lee predicted that someday soon, documents and data will not be merely displayed — but also processed, understood, and acted upon by both humans and machines (Berners-Lee et al., 2001).

Thirty years ago, Berners-Lee asked that *documents* be put on the web and linked using HTML (Hypertext Markup Language); now he asks that *data* be put on the web and linked using RDF (Resource Description Framework). His ask has become the rallying cry for the Semantic Web Community, who assert humanity's unadulterated right to all publicly available data sets, chanting:

"Raw Data Now!" *(TED Talks, 2009)*

To realize the potential of the semantic web, data will need to become more than open; it will need to become 5-star Open Data.



5 ★ OPEN DATA

★ make your stuff available on the Web (whatever format) under an open license[1]

★★ make it available as structured data (e.g., Excel instead of image scan of a table)[2]

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)[3]

★★★★ use URIs to denote things, so that people can point at your stuff[4]

★★★★★ link your data to other data to provide context[5]

Figure: The 5-star data rating scheme (Image: Hausenblas, 2012)

According to Berners-Lee, unlike documents on the web which are rendered by a web browser using HTML (Hypertext Markup Language), data on the web should be described using the Resource Description Framework (RDF). RDF is a data markup language whose syntax is expressed in a subject–verb–object linguistic pattern, one atomic unit of data in RDF is called a statement, also referred to as a triple, and follows this known language typology:

RDF Syntax
<subject> <predicate> <object>

Example Triple
<Margaret Mead> <employer> <American Museum of Natural History>

The *predicate* connects the *subject* and the *object* to form a *graph* that can be visualized as a network of nodes. Each component of the triple is represented by a resolvable uniform resource identifier (URI). These URIs allow data to be interlinked across the web, thereby eliminating silos and connecting globally distributed data stores.

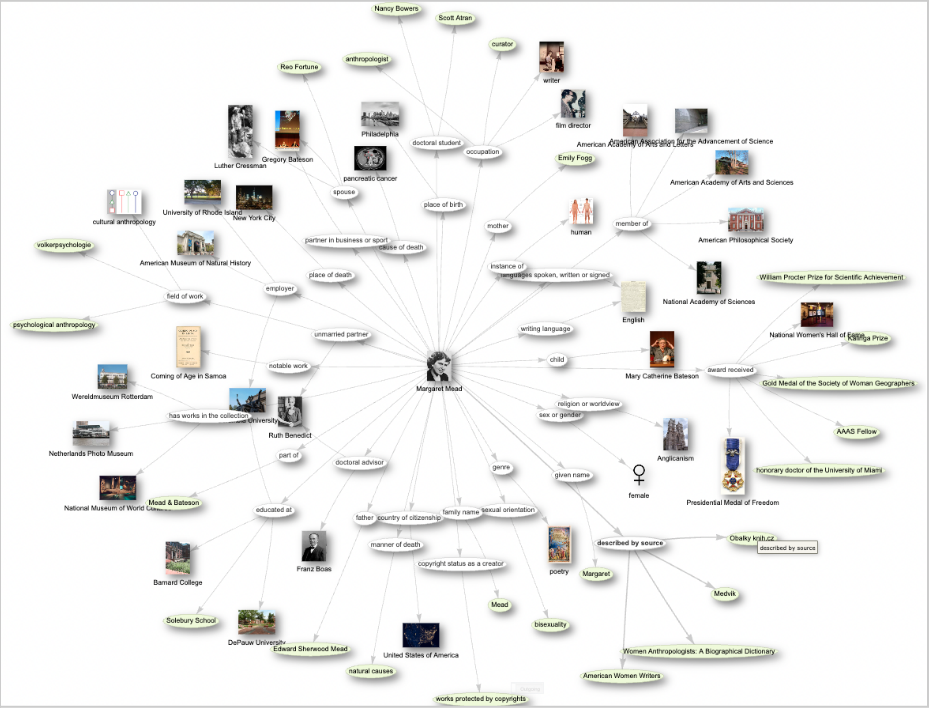


Figure: Margaret Mead's Wikidata knowledge graph (Image: Dearborn, 2022)

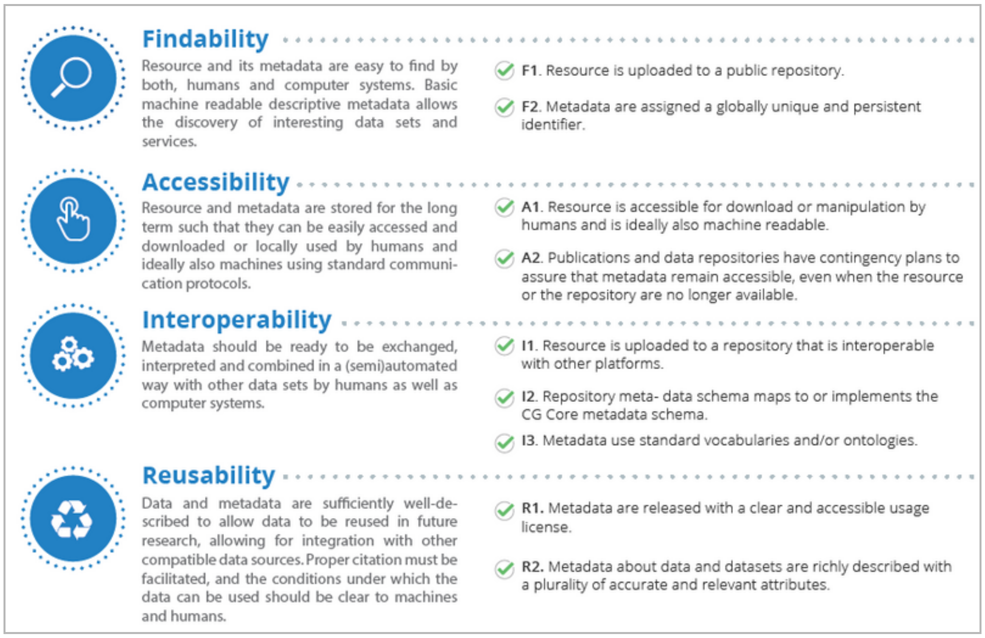
According to semantic web visionaries like Berners-Lee, if data practitioners describe data in RDF and abide by the 5-star data scheme, the world's dark data will come into the light, data silos will evaporate, and a new era for human knowledge will dawn.

FAIR Data Principles

Written in response to the dark data problem, specifically regarding the sciences, the FAIR data initiative is aimed at ensuring scientific reproducibility and society's ability to derive maximum benefit from public research investments. Naturally, humans are important stakeholders that stand to gain a lot from FAIR data. However, machines are increasingly the primary target of “FAIRified” data:

> *“…‘computational stakeholders’ are increasingly relevant, and demand as much, or more, attention as their importance grows. One of the grand challenges of data-intensive science, therefore, is to improve knowledge discovery through assisting both humans, and their computational agents, in the discovery of, access to, and integration and analysis of, task-appropriate scientific data and other scholarly digital objects” (Wilkinson, 2016).*

The scientific community has been working to better steward scholarly data and make science reproducible. Published in 2016, “The FAIR Guiding Principles for scientific data management and stewardship” have now been widely endorsed around the world (Wilkinson, 2016). FAIR is a mnemonic acronym that stands for its four guiding principles:



Findability

Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.

✓ **F1**. Resource is uploaded to a public repository.

✓ **F2**. Metadata are assigned a globally unique and persistent identifier.

Accessibility

Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.

✓ **A1**. Resource is accessible for download or manipulation by humans and is ideally also machine readable.

✓ **A2**. Publications and data repositories have contingency plans to assure that metadata remain accessible, even when the resource or the repository are no longer available.

Interoperability

Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.

✓ **I1**. Resource is uploaded to a repository that is interoperable with other platforms.

✓ **I2**. Repository meta- data schema maps to or implements the CG Core metadata schema.

✓ **I3**. Metadata use standard vocabularies and/or ontologies.

Reusability

Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.

✓ **R1.** Metadata are released with a clear and accessible usage license.

✓ **R2.** Metadata about data and datasets are richly described with a plurality of accurate and relevant attributes.

Figure: Checklist for FAIR organizational compliance (image: Open Access and Fair Principles, n.d.)

Bibliographic Framework (BIBFRAME)

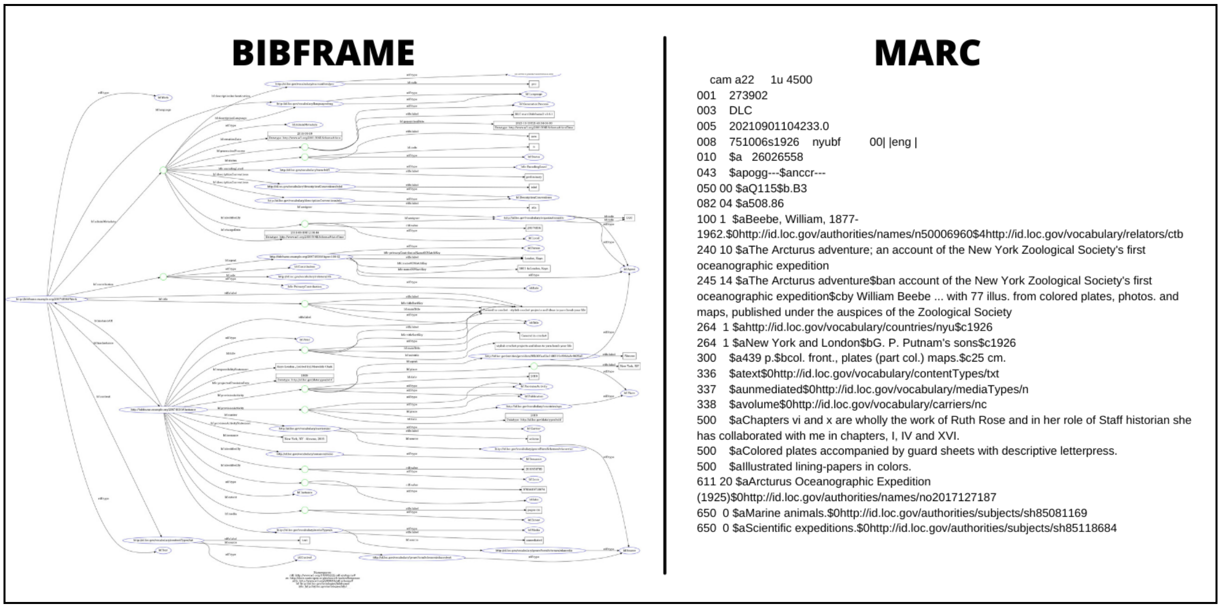
While the scientific community has been focused on making science reproducible, open, and FAIR, libraries have taken a brass-tacks approach to dark data. A decade after Berners-Lee presented his new vision for the web to the world, the Library of Congress and semantic consulting firm Zepheira co-developed an RDF ontology called Bibliographic Framework Initiative (BIBFRAME) with the goal of creating a bibliographic standard expressed in the W3C RDF standard that would replace the current cataloging standard, MARC (Machine Readable Cataloging Format) (Miller et al., 2012).

Long-cherished for its granularity and stability, MARC has served as the de-facto bibliographic data standard since the 1960s. Its success has made library cataloging a collaborative, efficient, global venture. Nevertheless, modern librarians lament that MARC is not web-friendly. Roy Tennant, the author of the now-infamous article "MARC Must Die" writes:

> *"Libraries exist to serve the present and future needs of a community of users. To do this well, they need to use the very best that technology has to offer. With the advent of the web, XML, portable computing, and other technological advances, libraries can become flexible, responsive organizations that serve their users in exciting new ways. Or not. If libraries cling to outdated standards, they will find it increasingly difficult to serve their clients as they expect and deserve." (Tennant, 2021).*

When the first version of BIBFRAME was completed in the Fall of 2012, the new data format was introduced as "the foundation for the future of bibliographic description that happens on, in, and as part of the web and the networked world we live in" (Miller et al., 2012). To facilitate the conversion process from MARC to RDF, the Library of Congress has created many conversion tools to assist organizations with their data transformation needs (Library of Congress, n.d.).

BIBFRAME is informed by the FRBR model (Functional Requirements for Bibliographic Records) and maintains the granularity of MARC. The standard is expressed in RDF's expected triple format and the vocabulary consists of three core classes: work, instance, and item (Schreur, 2018).



```
   cam a22    1u 4500
001   273902
003   DLC
005   20210901104233.0
008   751006s1926   nyubf       00| |eng |
010   $a   26026558
043   $apogg---$anccr---
050 00 $aQ115$b.B3
082 04 $a508.86
100 1  $aBeebe, William, 1877-
1962.$0http://id.loc.gov/authorities/names/n50006960$4http://id.loc.gov/vocabulary/relators/ctb
240 10 $aThe Arcturus adventure; an account of the New York Zoological Society's first
oceanographic expedition
245 14 $aThe Arcturus adventure$ban account of the New York Zoological Society's first
oceanographic expedition$cby William Beebe ... with 77 illus. from colored plates, photos. and
maps, published under the auspices of the Zoological Society
264  1 $ahttp://id.loc.gov/vocabulary/countries/nyu$c1926
264  1 $aNew York and London$bG. P. Putnam's sons$c1926
300   $a439 p.$bcol. front., plates (part col.) maps.$c25 cm.
336   $atext$0http://id.loc.gov/vocabulary/contentTypes/txt
337   $aunmediated$0http://id.loc.gov/vocabulary/mediaTypes/n
338   $avolume$0http://id.loc.gov/vocabulary/carriers/nc
500   $aChapters vi and x are wholly the work of Ruth Rose and in her role of Staff historian she
has collaborated with me in chapters, I, IV and XVI.
500   $aColored plates accompanied by guard sheets with descriptive letterpress.
500   $aIllustrated lining-papers in colors.
611 20 $aArcturus Oceanographic Expedition
(1925)$0http://id.loc.gov/authorities/names/no2017127187
650  0 $aMarine animals.$0http://id.loc.gov/authorities/subjects/sh85081169
650  0 $aScientific expeditions.$0http://id.loc.gov/authorities/subjects/sh85118684
```

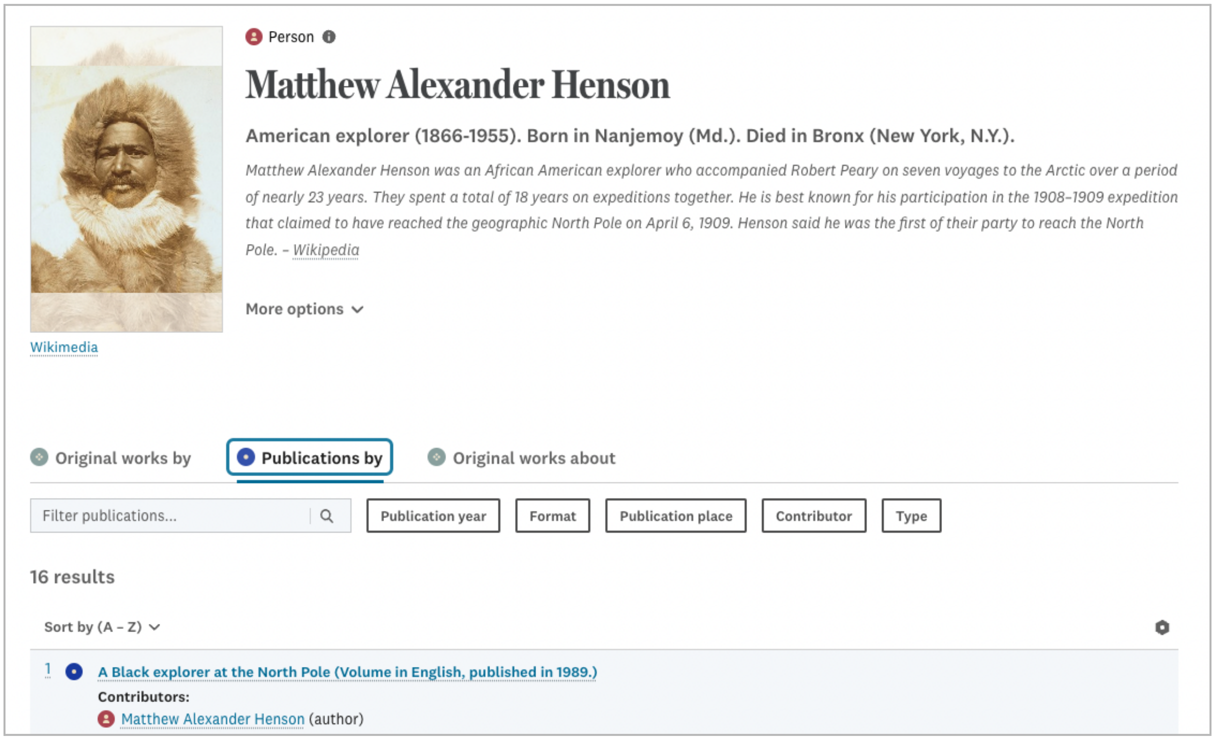
Figure: A bibliographic record for The Arcturus Adventure is expressed in BIBFRAME as a knowledge graph (left); the same bibliographic record is expressed in MARC (right). Which one is 5-star data? (Image: Dearborn, 2022).

Adoption of BIBFRAME has gained recent momentum with a 1.5 million dollar grant awarded by the Andrew W. Mellon Foundation to fund *Linked Data for Libraries*, dubbed *LD4L Labs* (2016-2018), distributed to Columbia, Cornell, Harvard, Library of Congress, Princeton, and Stanford. These six institutions are leading the charge by

- developing standards, guidelines, and infrastructure to communally produce metadata as linked open data;
- developing end-to-end workflows to create linked data in a technical services production environment;
- extending the BIBFRAME ontology to describe library resources in specialized domains and formats; and
- engaging the broader library community to ensure a sustainable and extensible environment (LD4P Partners, 2016).

To achieve these ambitious goals, the LD4P partners have organized a global community of practice around linked open data. Wikidata is playing a key role by providing librarians with an open, free, collaborative testing ground for semantic experimentation. The next two phases of the grant, Linked Data for Production: Pathway to Implementation (LD4P2) and Linked Data for Production: Closing the Loop (LD4P3) aim to mainstream BIBFRAME production with Sinopia, a global linked data creation environment (Stanford University Libraries, 2018).

Working in close collaboration with LD4P, Casalini Libri, an Italian library services firm, is converting entire institutional catalogs from MARC to BIBFRAME through their ShareVDE conversion service. Additionally, known named entities are being harmonized across catalogs with the assignment of globally persistent identifiers, called Supra Uniform Resource Identifiers (SFIDs.)



Person

Matthew Alexander Henson

American explorer (1866-1955). Born in Nanjemoy (Md.). Died in Bronx (New York, N.Y.).

Matthew Alexander Henson was an African American explorer who accompanied Robert Peary on seven voyages to the Arctic over a period of nearly 23 years. They spent a total of 18 years on expeditions together. He is best known for his participation in the 1908–1909 expedition that claimed to have reached the geographic North Pole on April 6, 1909. Henson said he was the first of their party to reach the North Pole. – Wikipedia

More options

Wikimedia

Original works by | Publications by | Original works about

Filter publications... | Publication year | Format | Publication place | Contributor | Type

16 results

Sort by (A – Z)

1 A Black explorer at the North Pole (Volume in English, published in 1989.)
Contributors:
Matthew Alexander Henson (author)

Figure: Matthew Alexander Henson, an African American Arctic Explorer's SUID is https://svde.org/agents/951654264154896. (Image: SVDE, 2023)

As of September 2021, the new version Share-VDE 2.0 beta went live with plans to bring the system into production in May 2023. The beta site already allows users to search nine catalogs in six languages, across 35,702,678 works. The user interface features accessible themes for vision-impaired persons, custom domain names, and branded skins for member institutions. (Share-VDE, 2021) Converted records are delivered back to adopting institutions as triple-store databases and delta loads of records from participating institutions are slated to occur in 2023.

At the Smithsonian Institution, the Share-VDE BIBFRAME conversion effort is being led by Descriptive Data Manager, Jackie Shieh, who is a huge linked data advocate and has worked with the Casalini Libri Team to usher nearly 2 million records from the Smithsonian Libraries and Archives catalog into the semantic web. In a recent blog, she writes:

> *"SVDE's back-end technology prepared the Smithsonian Libraries and Archives data for indexing, clustering, searching and representation. Their work helps data providers like the Smithsonian reap the benefits of linked data and connect library collections on the web in a user-friendly and informative manner." (Shieh, 2022)*

As the library world moves headlong towards BIBFRAME, it would be wise for BHL and its partners to proactively follow suit. One catalog at a time, Casalini Libri is helping libraries break free from the shackles of MARC and de-silo library data across the globe with a centralized semantic portal, Share-VDE.

The BIBFRAME initiative, FAIR principles, and 5-star linked open data have the potential to launch a gestalt leap for information, and ultimately human knowledge. These specifications are about harnessing, using, and stewarding data in ways that allow humans to find meaning in the zettabytes of noise.

1.3 BHL Data Management Today

Since its founding in 2006, the Biodiversity Heritage Library has provided free and open access to biodiversity information. Against a backdrop of institutional capacity constraints, BHL's global community of committed partners has built the world's largest open access digital biodiversity library. For over 17 years, BHL partners have been actively digitizing content and opening up access to historical literature and archives from natural history, botanical, research, and national libraries, as well as field naturalist clubs and scholarly societies. 182,000+ titles comprising over 60 million pages have been liberated from the physical shelves and rare book vaults of 549 contributing organizations. Complementing this global digitization operation, technical milestones have been centered on serving core stakeholders and expanding access to BHL's visual media, text, and data.

But the knowledge about Earth's biodiversity extends beyond institutional walls and stakeholders. The dissemination of data contained within BHL's collection and its integration into the semantic web reinforces BHL's commitment to repatriating biodiversity information, particularly to the local communities where the material was originally sourced.

1.4 BHL's Big Data Challenges

As strong as the BHL community's shared vision has proven to be, the backlog of user requests, metadata curation work, and the ever-present digitization request queue now represent many lifetimes of work for a disproportionately small group of engaged BHL staff. To rise to the looming challenges presented by climate change, BHL's data management strategies will need to pivot to embrace automation, crowdsourcing, machine learning, and the adoption of emerging semantic web standards. A recent review of BHL's information architecture by consulting firm Index Data yielded this final prescient insight:

> *"Aggregating biodiversity information is too big a job to be left to a relatively small cadre of information professionals." (Taylor et al., 2019).*

Today, BHL faces three big data challenges that must be solved to make the data in its corpus truly open, actionable, FAIR, and 5-star.

1. Correcting and Transcribing OCR Text Files

When a book is digitized, an unstructured text file is created by an Optical Character Recognition (OCR) engine. This unstructured text file is created alongside the page image and metadata files. Currently, BHL's OCR corpus is sizable: 289,000+ digital objects, comprising 60 million+ pages, amounting to 40+ gigabytes of data, silently awaiting conversion, normalization, and crowdsourcing.

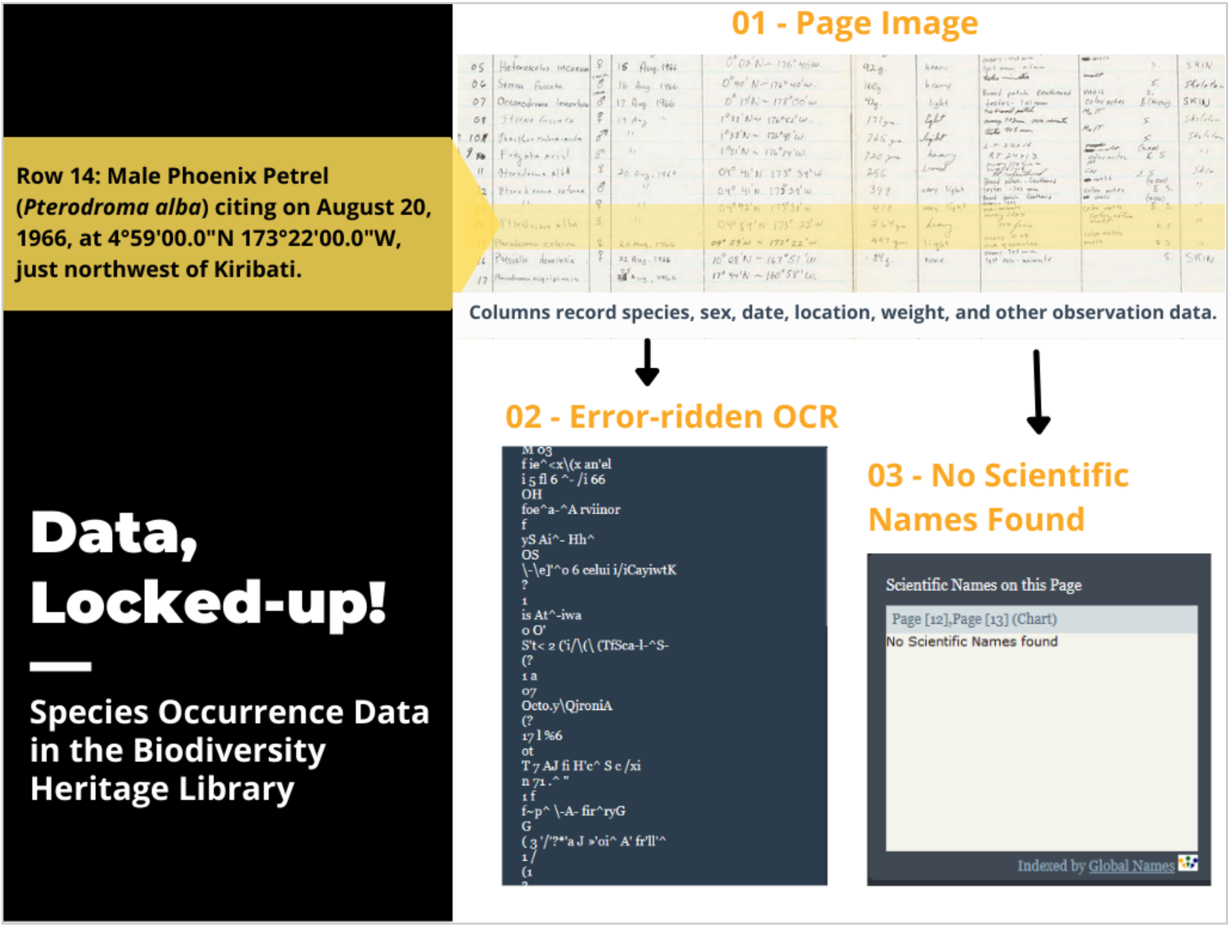
According to the 5-star rating scheme (Hausenblas, 2012), unstructured OCR text only ranks as 2-star data:

- It is not machine-readable;
- It does not contain URIs;
- It does not link to anything; and
- It is error-ridden.

BHL's Technical Coordinator and Lead Developer and the BHL Transcription Upload Tool Working Group (TUTWG) are working together to improve the quality of OCR text in the BHL corpus by

- reprocessing the corpus with Tesseract, an open-source OCR engine;
- experimenting with cutting-edge handwritten text recognition (HTR) engines for handwritten materials; and
- analyzing transcription platforms for their ability to extract data while hastening partner transcription initiatives.

Nevertheless, a sustainable, scalable workflow to liberate machine-readable data locked in BHL's OCR text files has yet to be forged.



Data, Locked-up!

Species Occurrence Data in the Biodiversity Heritage Library

Row 14: Male Phoenix Petrel (*Pterodroma alba*) citing on August 20, 1966, at 4°59'00.0"N 173°22'00.0"W, just northwest of Kiribati.

01 - Page Image

Columns record species, sex, date, location, weight, and other observation data.

02 - Error-ridden OCR

```
M 03
f ie^<x\(x an'el
i 5 fl 6 ^- /i 66
OH
foe^a-^A rviinor
f
yS Ai^- Hh^
OS
\-\e]'^o 6 celui i/iCayiwtK
?
1
is At^-iwa
o O'
S't< 2 ('i/\(\ (TfSca-l-^S-
(?
1 a
07
Octo.y\QjroniA
(?
17 l %6
ot
T 7 AJ fi H'c^ S c /xi
n 71 .^ "
1 f
f~p^ \-A- fir^ryG
G
( 3 '/'?*'a J »'oi^ A' fr'll'^
1 /
(1
```

03 - No Scientific Names Found

Scientific Names on this Page

Page [12],Page [13] (Chart)

No Scientific Names found

Indexed by Global Names

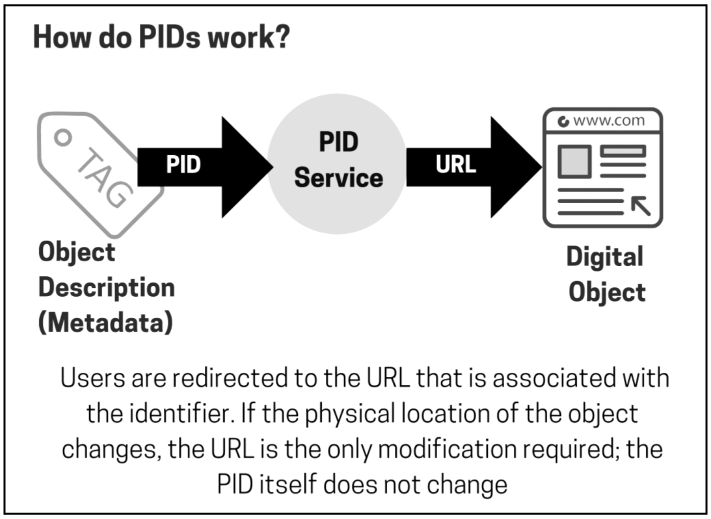
Figure: BHL's uncorrected OCR, particularly in data-rich archival materials, is "dark data." (Dearborn & Kalfatovic, 2022)

2. Improving BHL Search Precision and Retrieval

BHL users have asked for many search enhancements that require additional metadata which does not exist. Library resource description is constrained to collection, title, or item-level metadata for books, journals, and archival materials. More granular level record types, frequently referred to as "named entities,” such as articles, species, images, events, locations, nomenclatural acts, taxons, authors, and publishers, are described to a lesser extent. BHL’s suite of technical features like full-text search, taxonomic intelligence, and data model upgrades for scholarly articles, have improved access to previously uncatalogued content. Nevertheless, search functionality for under-described entities means a plethora of unique information in BHL’s collection is still quite difficult to retrieve. To make all of BHL’s content discoverable and reusable — beyond books and journals — programmatic metadata clean-up and enrichment, must become the strategic focus.

3. Linking and Depositing BHL Data with Global Knowledge Bases

The final tenet of the 5-star data scheme asks that “you link your data to other data to provide context.” Findings from BHL’s 2017 User Needs Assessment conducted by National Digital Stewardship Resident (NDSR), Pamela McClanahan found exactly this. “Linking” was consistently cited as a top BHL user request (McClanahan, 2018). For this reason, BHL’s Persistent Identifier Working Group (BHL-PIWG) has been actively registering DOIs (Digital Object Identifiers) to journal articles in BHL thereby bringing this content into the modern linked (5-star) network of scholarly research.



Users are redirected to the URL that is associated with the identifier. If the physical location of the object changes, the URL is the only modification required; the PID itself does not change

Figure: How do PIDs work? (Image: UCSB, 2020)

PIWG Chair and Manager at Biodiversity Heritage Library Australia, Nicole Kearney explains the benefits of persistent identifiers further:

> *"In modern online publishing, PID assignment and linking happens at the point of publication: DOIs (Digital Object Identifiers) for publications, ORCIDs (Open Researcher and Contributor IDs) for people, and RORs (Research Organization Registry IDs) for organisations. The DOI system provided by Crossref (the DOI registration agency for scholarly content) delivers reciprocal citations, enabling convenient clicking from article to article, and citation tracking, enabling authors and institutions to track the impact and reach of their research output. Publications that lack PIDs, which include the vast majority of legacy literature, are hard to find and sit outside the linked network of scholarly research."*
> *(Kearney et al., 2021)*

Additionally, BHL's Cataloging and Metadata Committee is proactively adding persistent identifiers to BHL author records with a recent harvest of 88,000+ URIs from Wikidata in 2022 (Dearborn & Leachman, 2023). The hard work to interlink BHL's persistent identifiers with external authoritative identifiers is underway and a comprehensive URI policy for all entities in BHL is currently being drafted collaboratively by BHL working groups. Luckily, a powerful information broker has emerged to hasten BHL's progress: Wikidata.

2.0 Wikidata

2.1 About

At the heart of all this interlinking work is Wikidata, Wikimedia Foundation's most active project in terms of total edits and item count (2023, Appendix 3: Statistics). The platform is generating and providing 5-star linked open data to the world at break-neck speed (Redi, 2018). The idea for Wikidata was born in 2012 out of a desire to maintain synchrony between the Wikimedia Foundation's universe of wikis (300+

language editions of Wikipedia and sister Wikimedia projects). The time-consuming task of manually updating articles in each language edition whenever a statistic, figure, or fact changed about the world was inherently unsustainable.

Dr. Denny Vrandečić and Dr. Markus Krötzsch were tasked in 2012 to lead a team of Wikimedia Deutschland developers to build Wikidata (Perez, 2012). Today, the collaboratively edited knowledge base has grown to be the Internet's largest crowdsourced linked open data repository (Haller et al., 2022). The Wikidata community has over 24,000+ active users who have edited 102M+ items over 1,800,000,000 times (Appendix: Statistics). All data in Wikidata is licensed under Creative Commons CC0 License, free to all.

> *"Wikidata helps with the problem of opening data silos. There is a lot of knowledge out there, we know a lot about our surroundings, but it is hidden behind paywalls in silos. Wikidata bursts these silos open for the greater public to be used." –Andra Waagmeester, Micelio Bioinformatician. (Wikimedia Deutschland, 2019)(Appendix 2:Interviews).*

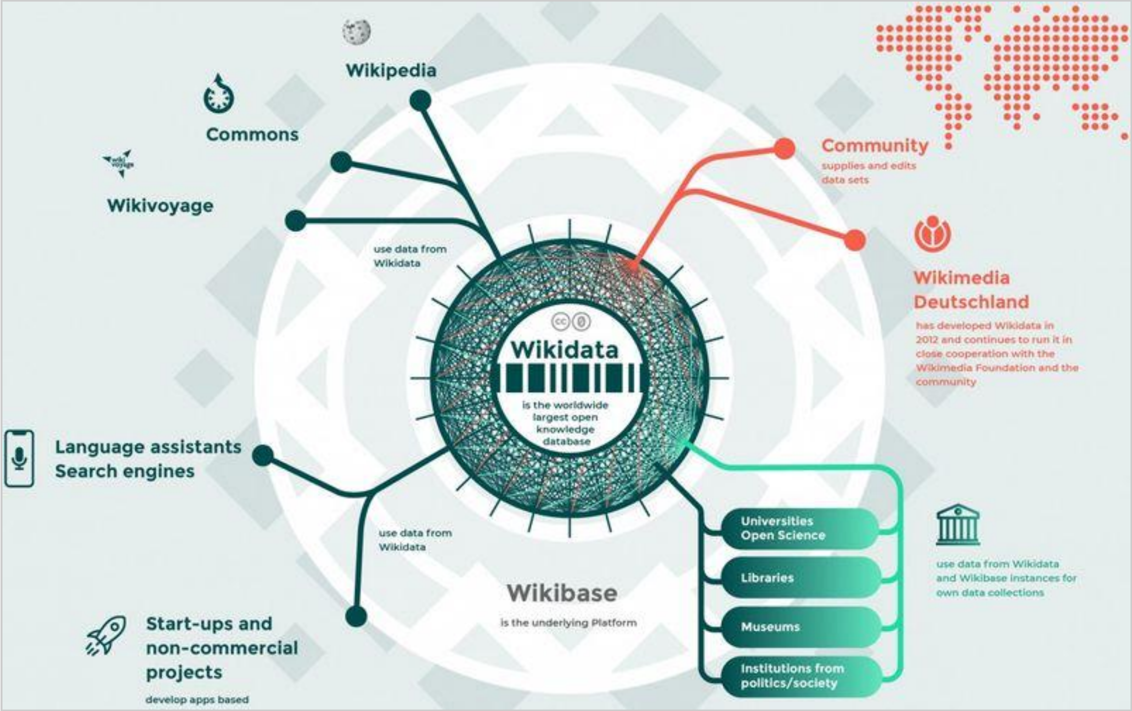

Wikipedia
Commons
Wikivoyage
use data from Wikidata
Community
supplies and edits data sets
Wikimedia Deutschland
has developed Wikidata in 2012 and continues to run it in close cooperation with the Wikimedia Foundation and the community
Wikidata
is the worldwide largest open knowledge database
Language assistants
Search engines
use data from Wikidata
Wikibase
is the underlying Platform
Universities Open Science
Libraries
Museums
Institutions from politics/society
use data from Wikidata and Wikibase instances for own data collections
Start-ups and non-commercial projects
develop apps based

Figure: The Wikidata Ecosystem - How the free knowledge database Wikidata works Grafik by MOR for Wikimedia Deutschland. (Wikimedia Deutschland, 2019)

Wikidata 101

In Wikidata, a statement is the atomic unit of data, rendered in the RDF triple format. (Wikipedia Contributors, 2022)

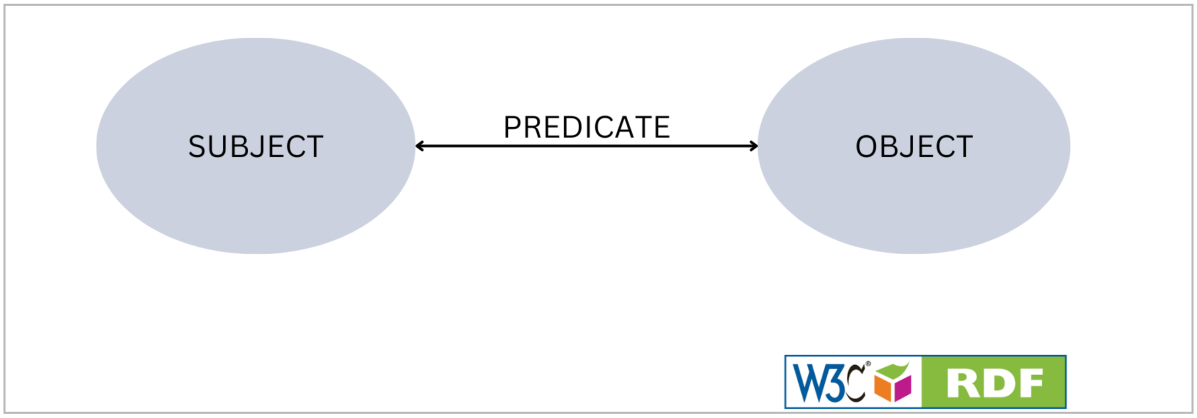


Figure: An RDF triple is an atomic unit of data (W3C, 2014); in Wikidata these are referred to as statements or claims. (Image: Dearborn, 2023)

Additionally, every Wikidata entity receives a unique identifier beginning with the prefix **Q** for items and **P** for properties. A persistent URL may be obtained for any item by appending the unique ID (such as Q11575 or P8724) to the Wikidata namespace: http://www.wikidata.org/entity/.

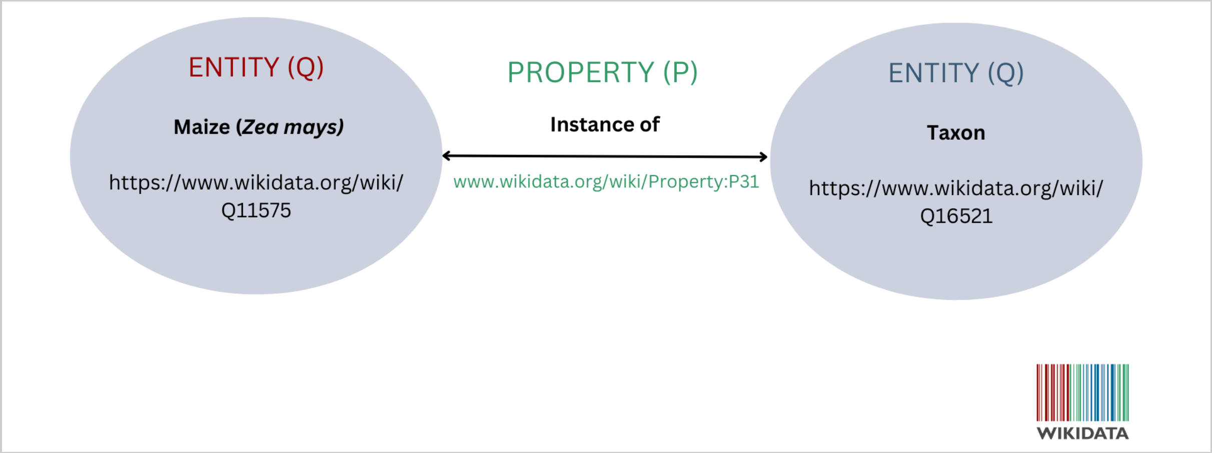
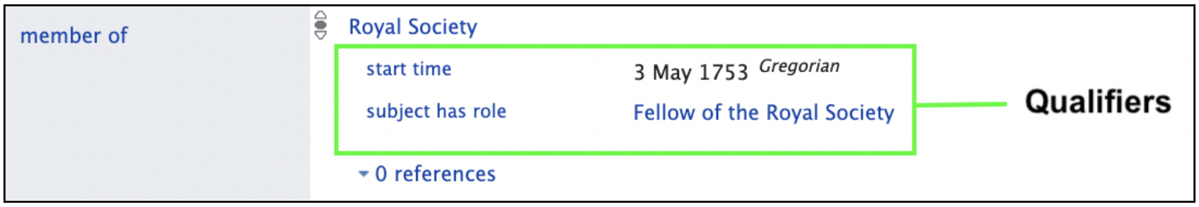

ENTITY (Q)
Maize (Zea mays)
https://www.wikidata.org/wiki/Q11575
PROPERTY (P)
Instance of
www.wikidata.org/wiki/Property:P31
ENTITY (Q)
Taxon
https://www.wikidata.org/wiki/Q16521
WIKIDATA

Figure: An RDF triple rendered in Wikidata as a statement: Entity (Q) - Property (P) - Entity (Q), connected with resolvable URIs. (Image: Dearborn, 2023)

Statements may also include qualifiers that expand the meaning beyond a simple property-value pair and "further specify the application of a property, to constrain the validity of the value, or to give additional details about the value.”



member of

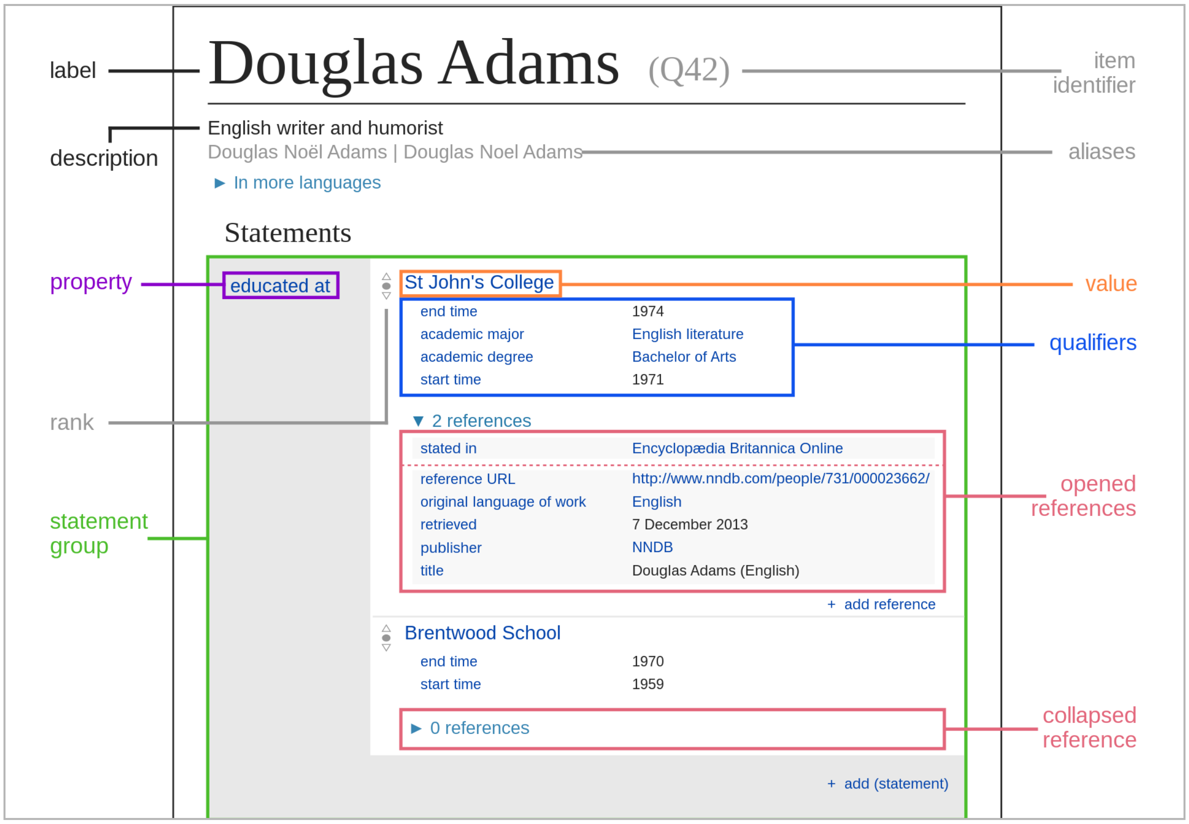
Royal Society

start time	3 May 1753 *Gregorian*
subject has role	Fellow of the Royal Society

▾ 0 references

Figure: Qualifiers further clarify a Wikidata statement about Carl Linnaeus (Q1043), Swedish botanist, physician, and zoologist (1707–1778), and his membership as a Fellow of the Royal Society. (Image: Wikidata Q1043)

Below is the anatomy of a Wikidata item and the general terminology associated with a Wikidata Q record.



Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

► In more languages

Statements

educated at

St John's College

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

+ add reference

Brentwood School

end time	1970
start time	1959

► 0 references

+ add (statement)

Figure: A graphic representation of the data model used for items. (Image: Wikimedia UX Designer, Charlie Kritschmar)

It is important to note that Wikidata statements are not facts, they are claims. These claims can be added by Wikimedians with zero to many references to further substantiate the claim. BHL Staff should be invested in curating statement references to make Wikidata more authoritative. For an in-depth overview of the Wikidata data model, consult the primer.

2.2 Identified Use Cases

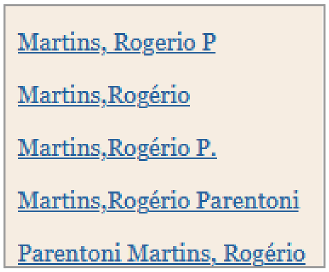
2.2.1 Name Disambiguation

The problem of identity in the historical record is a perennial issue for the scientific community. An article published in *Earth Science Informatics* assessing various identifier schemes reports:

> *"The problem of identity has vexed humanity throughout all of recorded history. A wide variety of methods; from assigned identifiers to taxonomic techniques and beyond; have historically been used to resolve the issue of whether this thing, whatever or whomever it may be, is what it purports to be" (Duerr et al., 2011).*

To help solve the problem, Wikidata proves immensely useful, acting as a powerful identifier broker and universally accessible global name authority registry. Wikidata's appeal over other authority services comes down to the registration process — it is easy, instant, and almost anyone can do it. This lack of gatekeeping challenges traditional hierarchical notions of "authority" as in typical Wikimedian spirit, the de-facto "authority" is always the semi-anonymous, self-correcting crowd.

Collecting identifiers linked to *BHL Creator IDs* is a tactic employed by BHL's Cataloging and Metadata Committee and the Tech Team. This strategy helps to eliminate duplicate authors in the BHL database.



Martins, Rogerio P

Martins,Rogério

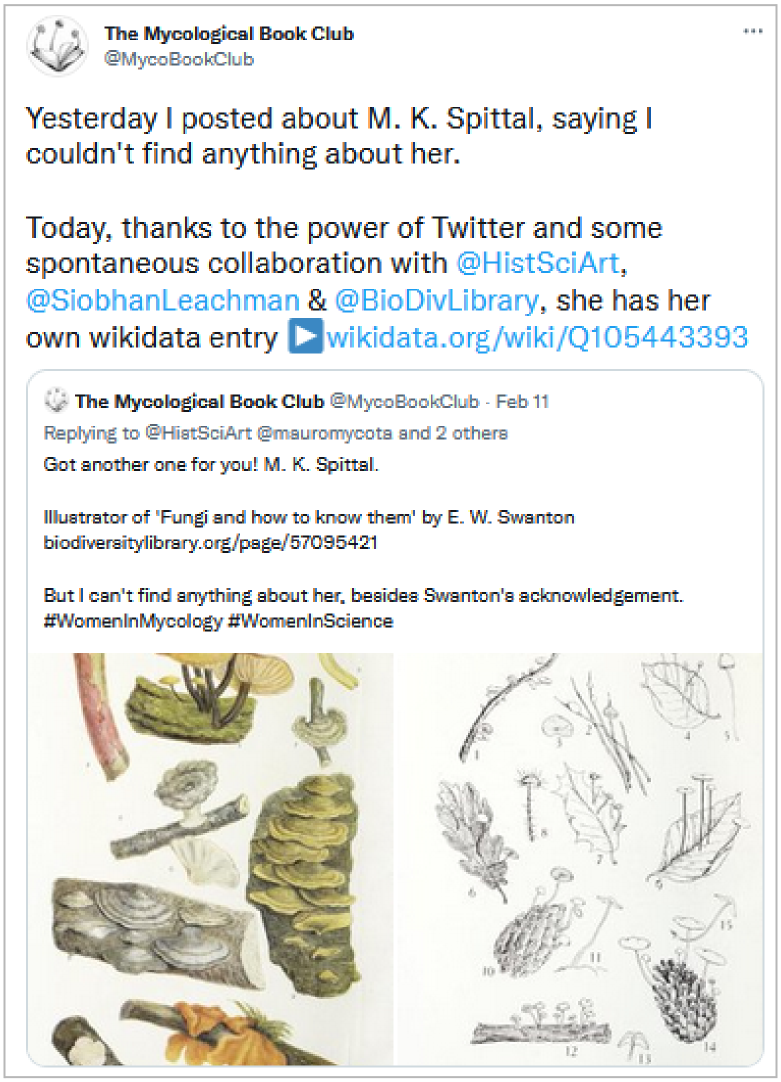
Martins,Rogério P.

Martins,Rogério Parentoni

Parentoni Martins, Rogério

Figure: Who's who? An example of multiple author names on the BHL website results from metadata aggregated from hundreds of contributors. (Image: Dearborn, 2022)

Curating and collecting persistent identifiers for people and organizations in BHL goes beyond disambiguation. Ample evidence that the curation (and proliferation) of *BHL Creator IDs,* uncover the contributions of under-represented groups such as women scientific illustrators like Mary K. Spittal (Duerr et al., 2011).



The Mycological Book Club
@MycoBookClub

Yesterday I posted about M. K. Spittal, saying I couldn't find anything about her.

Today, thanks to the power of Twitter and some spontaneous collaboration with @HistSciArt, @SiobhanLeachman & @BioDivLibrary, she has her own wikidata entry ▶wikidata.org/wiki/Q105443393

The Mycological Book Club @MycoBookClub · Feb 11
Replying to @HistSciArt @mauromycota and 2 others

Got another one for you! M. K. Spittal.

Illustrator of 'Fungi and how to know them' by E. W. Swanton
biodiversitylibrary.org/page/57095421

But I can't find anything about her, besides Swanton's acknowledgement.
#WomenInMycology #WomenInScience

Figure: Serendipitous interaction on Twitter surfaces Mary K. Spittal, scientific illustrator, in BHL and Wikidata. (Image: Marshall, 2021)

Mary K. Spittal is no longer obscured by missing data points. With a *BHL Creator ID* added to both BHL and Wikidata, those records have been interlinked, and the publications Spitall helped create now surface through a simple Google search. Spittal and her illustrations are now part of Wikidata's growing biodiversity knowledge graph.

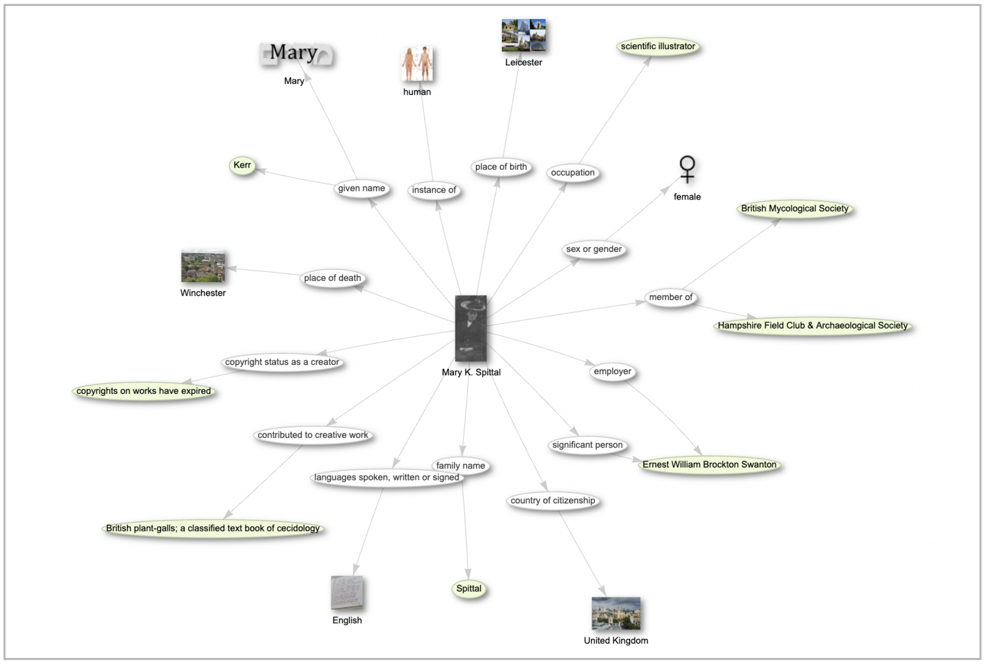

Mary
Mary
human
Leicester
scientific illustrator
Kerr
given name
instance of
place of birth
occupation
♀
female
British Mycological Society
sex or gender
Winchester
place of death
member of
Hampshire Field Club & Archaeological Society
copyright status as a creator
copyrights on works have expired
Mary K. Spittal
employer
contributed to creative work
significant person
Ernest William Brockton Swanton
family name
languages spoken, written or signed
country of citizenship
British plant-galls; a classified text book of cecidology
English
Spittal
United Kingdom

Figure: Mary K. Spittal, once an obscure data point, now has meaningful semantic interlinkages to other Wikidata entities. (Image: Dearborn, 2022 https://w.wiki/6S7V)

Despite many promising developments on the horizon for identity management, the un-disambiguated data that is in BHL must be dealt with now. BHL researchers continue to encounter information dead-ends and big data aggregators that consume BHL data for their discovery layers (OCLC, DPLA, CrossRef, OpenAlex) only replicate the same problem for their end-users.

In the recent paper “People are Essential to Linking Biodiversity Data,” written by BHL’s Persistent Identifier Working Group Chair, Nicole Kearney, and BHL advisors Dr. Rod Page and Siobhan Leachman et. al., argue that providing users with the basic ability to differentiate between people is fundamentally important:

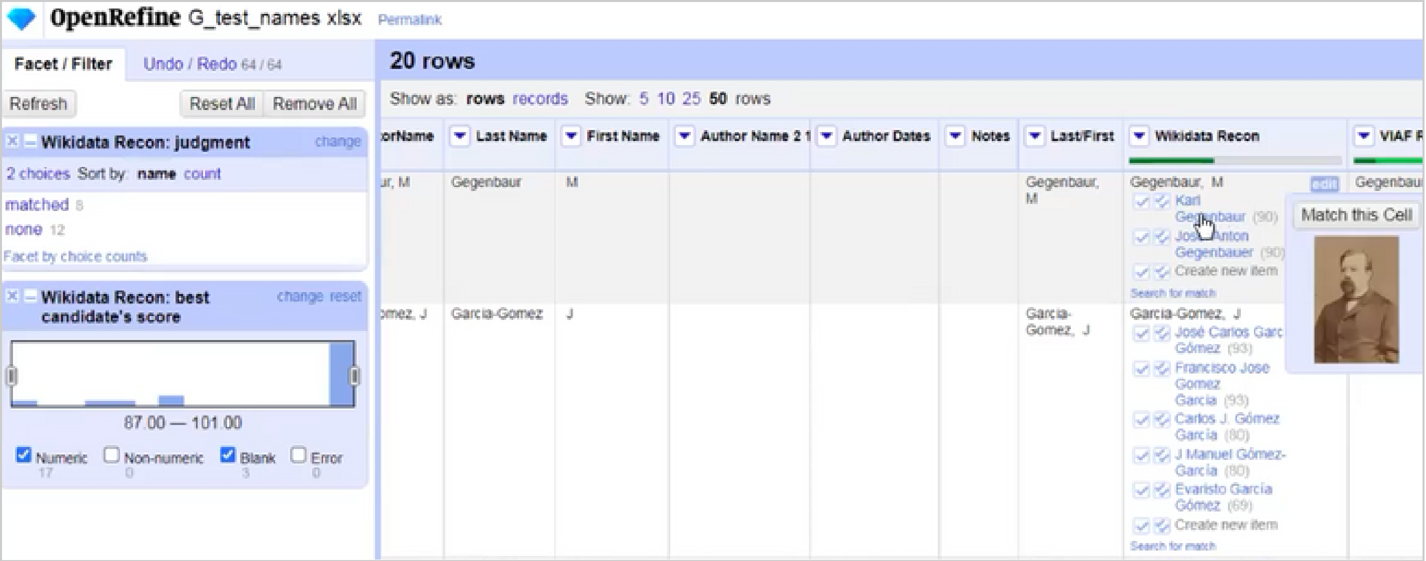
> *“Person data are almost always associated with entities such as specimens, molecular sequences, taxonomic names, observations, images, traits and publications [...] these entities are also useful in validating data, integrating data across collections and institutional databases and can be the basis of future research into biodiversity and science” (Groom et al., 2020).*

To help solve the tumult of duplicate author names after they have come into BHL *as name strings* from numerous sources, BHL's Cataloging and Metadata Committee has piloted three workflows, as part of the group's long-standing Author Merge project.

1. OpenRefine Wikidata Extension
2. Wikidata's Mix'N'Match Tool
3. Round Tripping Persistent Identifiers from Wikidata

Workflow 1 — OpenRefine Wikidata Extension

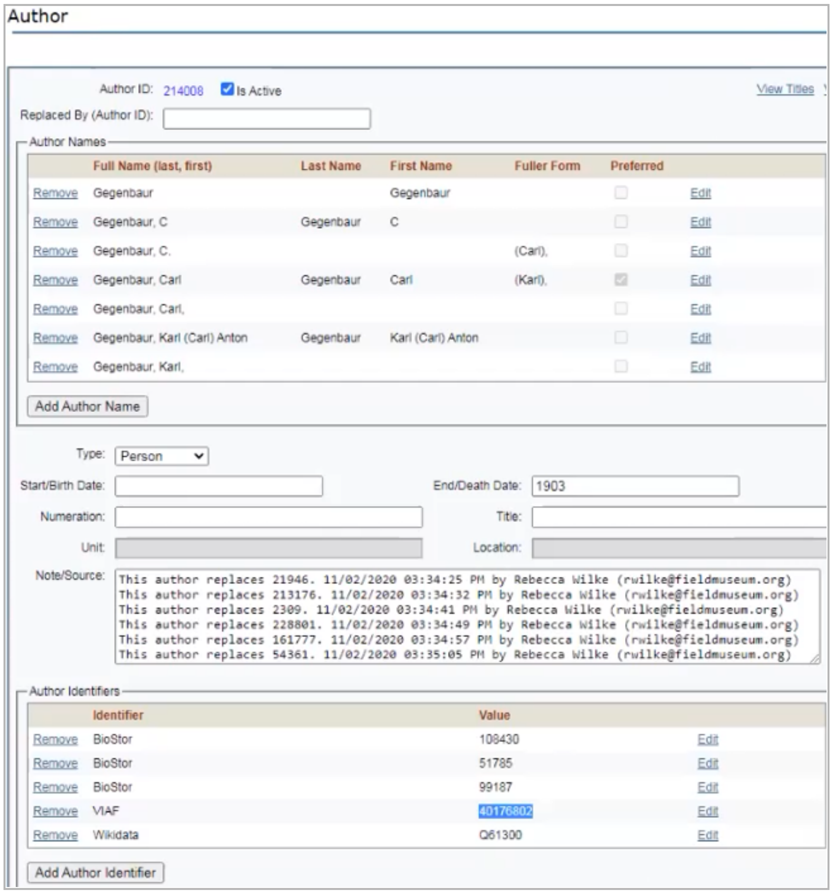
Diana Duncan, former BHL Cataloging and Metadata Chair and Lead Cataloger at the Field Museum, has documented an OpenRefine reconciliation workflow. The purpose is to match BHL free-text author name data with identifiers from external services like VIAF, ORCID, and Wikidata. (Appendix 2: Interviews) Of these services, Duncan likes the Wikidata extension best because it provides on-screen images, extra metadata points, and the ability to edit queries on the fly. These features help immensely to expedite the name reconciliation process.



orName	Last Name	First Name	Author Name 2 1	Author Dates	Notes	Last/First	Wikidata Recon	VIAF R
ur, M	Gegenbaur	M				Gegenbaur, M	Gegenbaur, M Karl Gegenbaur (90) José Anton Gegenbauer (90) Create new item Search for match	Gegenbau
omez, J	Garcia-Gomez	J				Garcia-Gomez, J	Garcia-Gomez, J José Carlos Garc Gómez (93) Francisco Jose Gomez Garcia (93) Carlos J. Gómez Garcia (80) J Manuel Gómez-Garcia (80) Evaristo Garcia Gómez (69) Create new item Search for match	

Figure: Reconciling BHL Author Names using OpenRefine and the Wikidata Reconciliation service (Image: Dearborn, 2021)

For each reconciled batch of roughly a thousand names, BHL staff merge records, add administrative notes, and the matched URIs to BHL's administrative back-end.



Author

Author ID: 214008 ☑ Is Active

Replaced By (Author ID):

Author Names

	Full Name (last, first)	Last Name	First Name	Fuller Form	Preferred	
Remove	Gegenbaur		Gegenbaur		□	Edit
Remove	Gegenbaur, C	Gegenbaur	C		□	Edit
Remove	Gegenbaur, C.			(Carl),	□	Edit
Remove	Gegenbaur, Carl	Gegenbaur	Carl	(Karl),	☑	Edit
Remove	Gegenbaur, Carl,				□	Edit
Remove	Gegenbaur, Karl (Carl) Anton	Gegenbaur	Karl (Carl) Anton		□	Edit
Remove	Gegenbaur, Karl,				□	Edit

Type: Person

Start/Birth Date:

End/Death Date: 1903

Numeration:

Title:

Unit:

Location:

Note/Source:

```
This author replaces 21946. 11/02/2020 03:34:25 PM by Rebecca Wilke (rwilke@fieldmuseum.org)
This author replaces 213176. 11/02/2020 03:34:32 PM by Rebecca Wilke (rwilke@fieldmuseum.org)
This author replaces 2309. 11/02/2020 03:34:41 PM by Rebecca Wilke (rwilke@fieldmuseum.org)
This author replaces 228801. 11/02/2020 03:34:49 PM by Rebecca Wilke (rwilke@fieldmuseum.org)
This author replaces 161777. 11/02/2020 03:34:57 PM by Rebecca Wilke (rwilke@fieldmuseum.org)
This author replaces 54361. 11/02/2020 03:35:05 PM by Rebecca Wilke (rwilke@fieldmuseum.org)
```

Author Identifiers

	Identifier	Value	
Remove	BioStor	108430	Edit
Remove	BioStor	51785	Edit
Remove	BioStor	99187	Edit
Remove	VIAF	40176802	Edit
Remove	Wikidata	Q61300	Edit

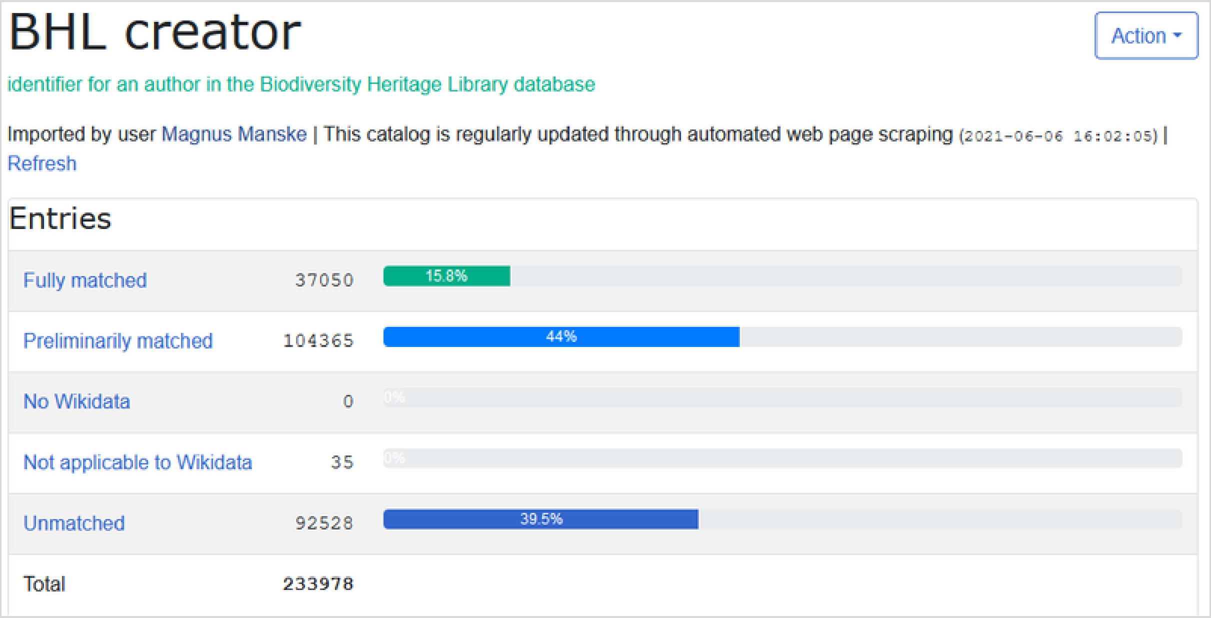
Figure: Merging authors in the BHL Administrative Dashboard, adding URIs and extra data points is a painstaking investigative and curation process for BHL's catalogers. (Image: Dearborn, 2021)

Workflow 2 - Wikidata's Mix'n'match Tool

Wikidata's Mix'n'match tool, created by veteran Wikimedian and Media Wiki Developer Magnus Manske, is providing a low-barrier way of interlinking *BHL Creator IDs* to corresponding Wikidata items.

Records from over 3,400 institutional datasets are being converted to linked open data and assigned identifiers by the Mix'n'match tool. Two of BHL's datasets are going semantic: BHL's *Creator IDs* and *Bibliography ID*s. Since July 2017, over 37,000 BHL *Creator IDs* have been processed.

Siobhan Leachman, a Wikimedian and active BHL volunteer, is a big advocate of the tool. Leachman has made over 15,000 author name matches alone! In addition to gamifying Wikidata statement creation, the tool supports catalog updates, status reports, and tracks user activity.



BHL creator

identifier for an author in the Biodiversity Heritage Library database

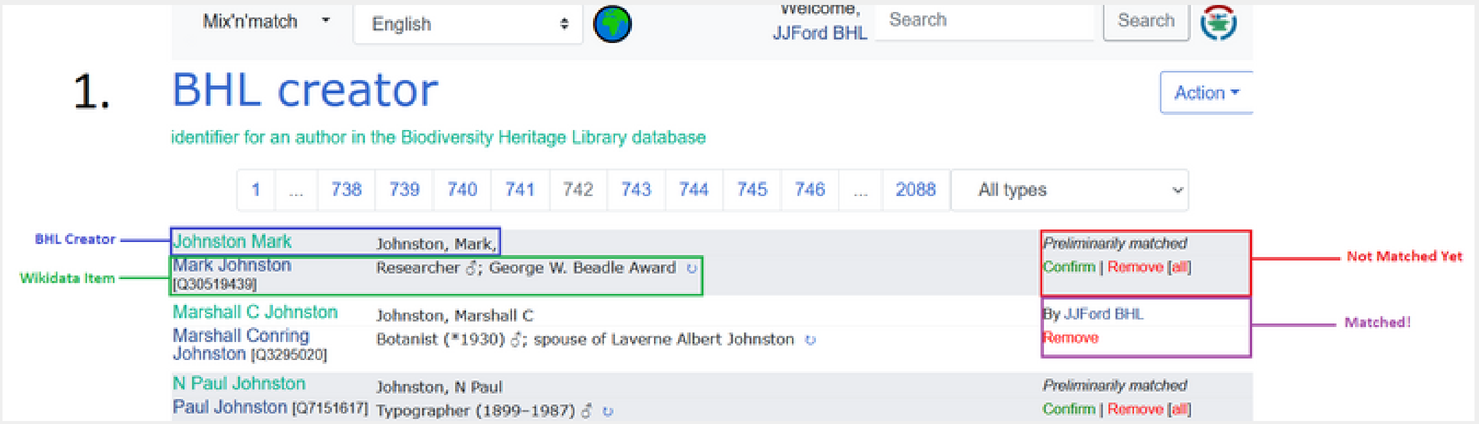
Imported by user Magnus Manske | This catalog is regularly updated through automated web page scraping (2021-06-06 16:02:05) | Refresh

Entries

Fully matched	37050	15.8%
Preliminarily matched	104365	44%
No Wikidata	0	0%
Not applicable to Wikidata	35	0%
Unmatched	92528	39.5%
Total	233978	

Figure: The Mix’n’match status report for BHL Creator IDs (Image: MixNMatch)

To make matches, the interface presents an auto-match to be inspected more closely by the player. Reconnaissance work is done to confirm the match.



1. BHL creator

identifier for an author in the Biodiversity Heritage Library database

Johnston Mark	Johnston, Mark,	*Preliminarily matched*
Mark Johnston [Q30519439]	Researcher ♂; George W. Beadle Award ↻	Confirm | Remove [all]
Marshall C Johnston	Johnston, Marshall C	By JJFord BHL
Marshall Conring Johnston [Q3295020]	Botanist (*1930) ♂; spouse of Laverne Albert Johnston ↻	Remove
N Paul Johnston	Johnston, N Paul	*Preliminarily matched*
Paul Johnston [Q7151617]	Typographer (1899–1987) ♂ ↻	Confirm | Remove [all]

Figure: The Mix’n’match interface (Image: Dearborn, 2022)

In a recent interview, Leachman endorses Mix’n’Match as *the* perfect entry point for BHL staff who have no experience with Wikidata: it’s easy, it’s fun, and the linking factor is incredibly powerful. ([Appendix 2: Interviews](#))

Workflow 3 - Round Tripping Persistent Identifiers from Wikidata

In a recent blog post, “Biodiversity Heritage Library is Round Tripping Persistent Identifiers with the Wikidata Query Service BHL” BHL’s Committee members documented their workflow that enabled BHL’s harvest of 88,507 author identifiers. (Dearborn & Leachman et.al., 2022)

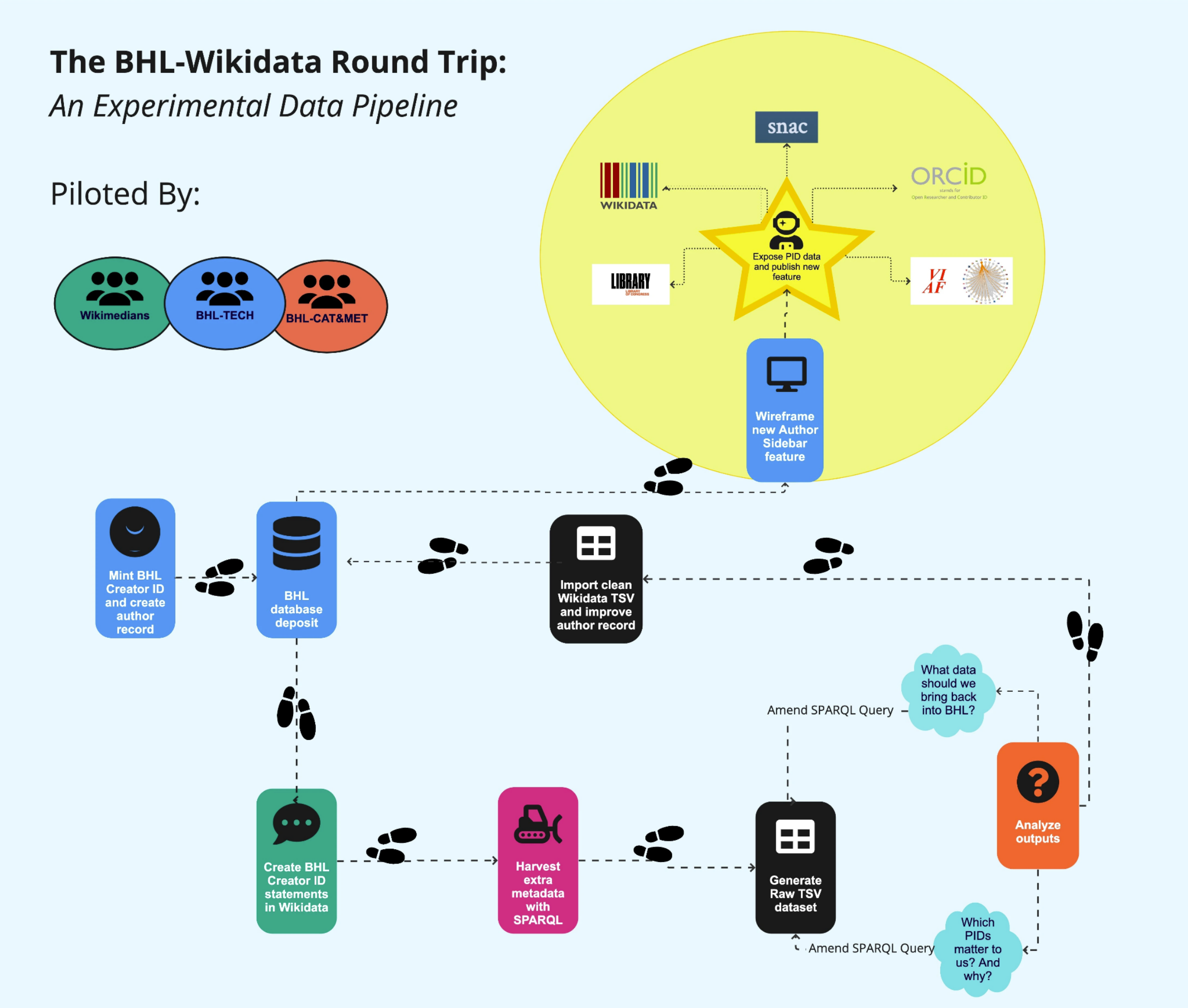
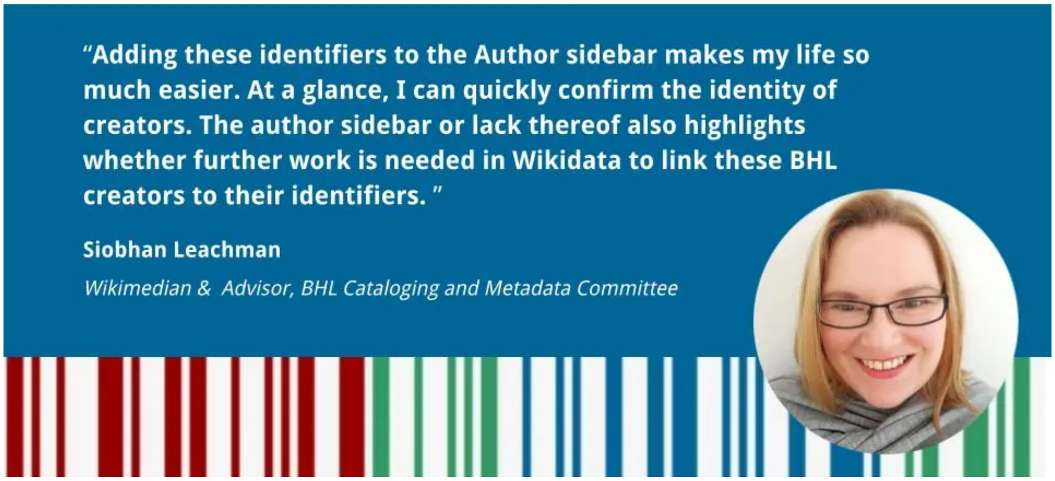

The BHL-Wikidata Round Trip:
An Experimental Data Pipeline
Piloted By:
Wikimedians
BHL-TECH
BHL-CAT&MET
snac
WIKIDATA
ORCID
LIBRARY
VIAF
Expose PID data and publish new feature
Wireframe new Author Sidebar feature
Mint BHL Creator ID and create author record
BHL database deposit
Import clean Wikidata TSV and improve author record
What data should we bring back into BHL?
Amend SPARQL Query
Analyze outputs
Create BHL Creator ID statements in Wikidata
Harvest extra metadata with SPARQL
Generate Raw TSV dataset
Amend SPARQL Query
Which PIDs matter to us? And why?

Figure: The BHL-Wikidata Round Trip, was an experimental data pipeline piloted by BHL Committee Members in Spring 2022, made possible by BHL Staff and Wikimedian curation efforts. (Image: Dearborn, 2021)

This workflow later resulted in the development of an author sidebar on the BHL platform to expose the author data in BHL’s database to users doing name research and disambiguation work. So far, the feedback has been overwhelmingly positive.



"Adding these identifiers to the Author sidebar makes my life so much easier. At a glance, I can quickly confirm the identity of creators. The author sidebar or lack thereof also highlights whether further work is needed in Wikidata to link these BHL creators to their identifiers. "

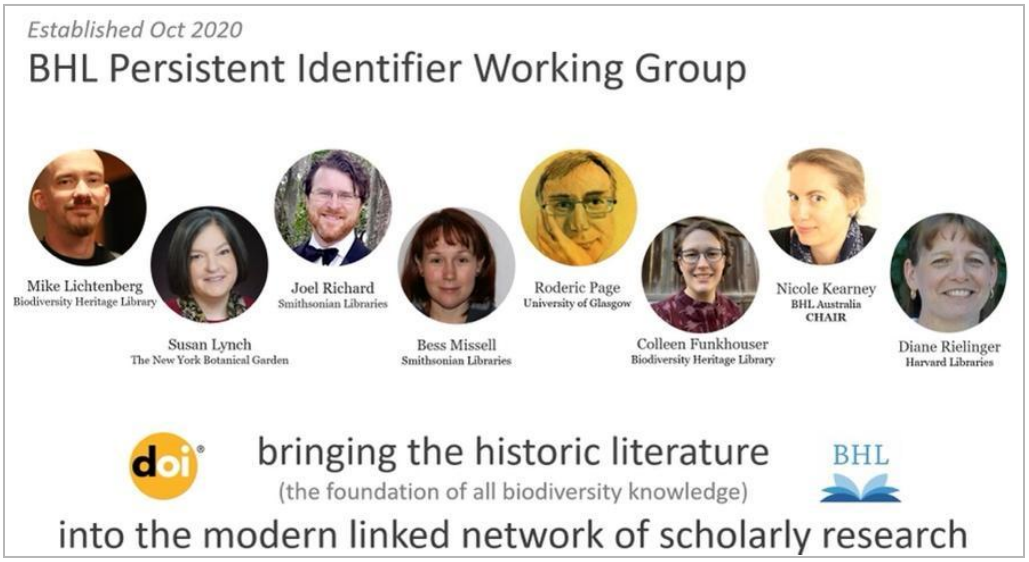
Siobhan Leachman

Wikimedian & Advisor, BHL Cataloging and Metadata Committee

Figure: BHL has added author data to BHL's front-end to help Wikimedians do their Wikidata work more effectively. (Image: Leachman & Dearborn, 2021)

2.2.2 Adding BHL Articles and DOIs to Wikidata

Established in 2020, the BHL Persistent Identifier Working Group (PIWG) has been generating article metadata for BHL and assigning Digital Object Identifiers (DOIs) to this previously undescribed, and thus unfindable, content.



Established Oct 2020

BHL Persistent Identifier Working Group

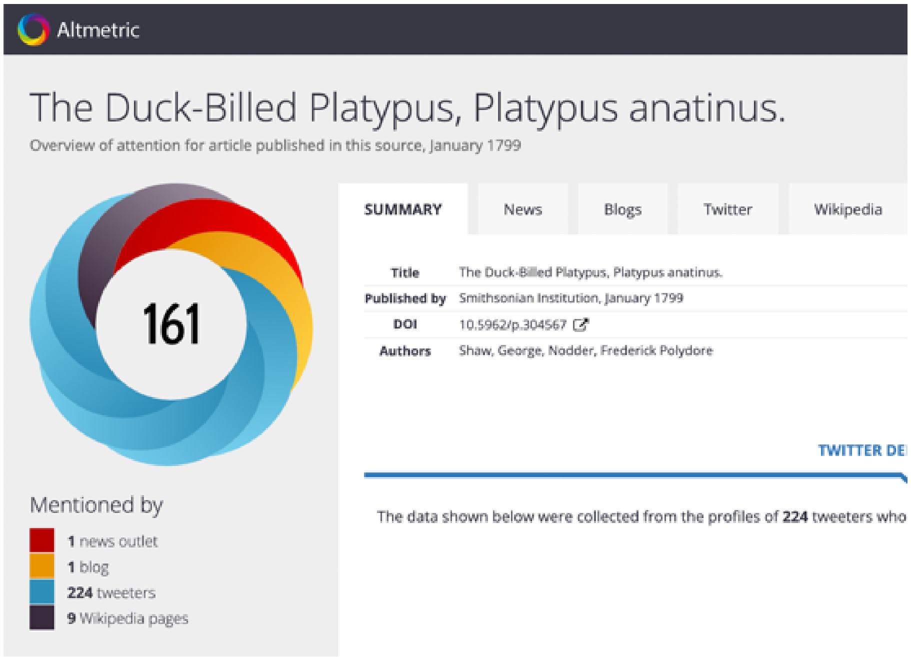
bringing the historic literature

(the foundation of all biodiversity knowledge)

into the modern linked network of scholarly research

Figure: PIWG is comprised of BHL staff committed to creating access points for articles in BHL, and more broadly the web. (Image: [What Is BHL's New Persistent Identifier Working Group DOI'ng?](#))

PIWG has developed best practices, workflows, tools, documentation, and videos to facilitate article metadata creation and DOI assignment for BHL partner institutions and their publications. The article creation process is work-intensive but incredibly valuable to the scholarly community and also helps to track the usage of BHL content across the web.



Altmetric

The Duck-Billed Platypus, Platypus anatinus.

Overview of attention for article published in this source, January 1799

Title	The Duck-Billed Platypus, Platypus anatinus.
Published by	Smithsonian Institution, January 1799
DOI	10.5962/p.304567
Authors	Shaw, George, Nodder, Frederick Polydore

TWITTER DE

The data shown below were collected from the profiles of 224 tweeters who

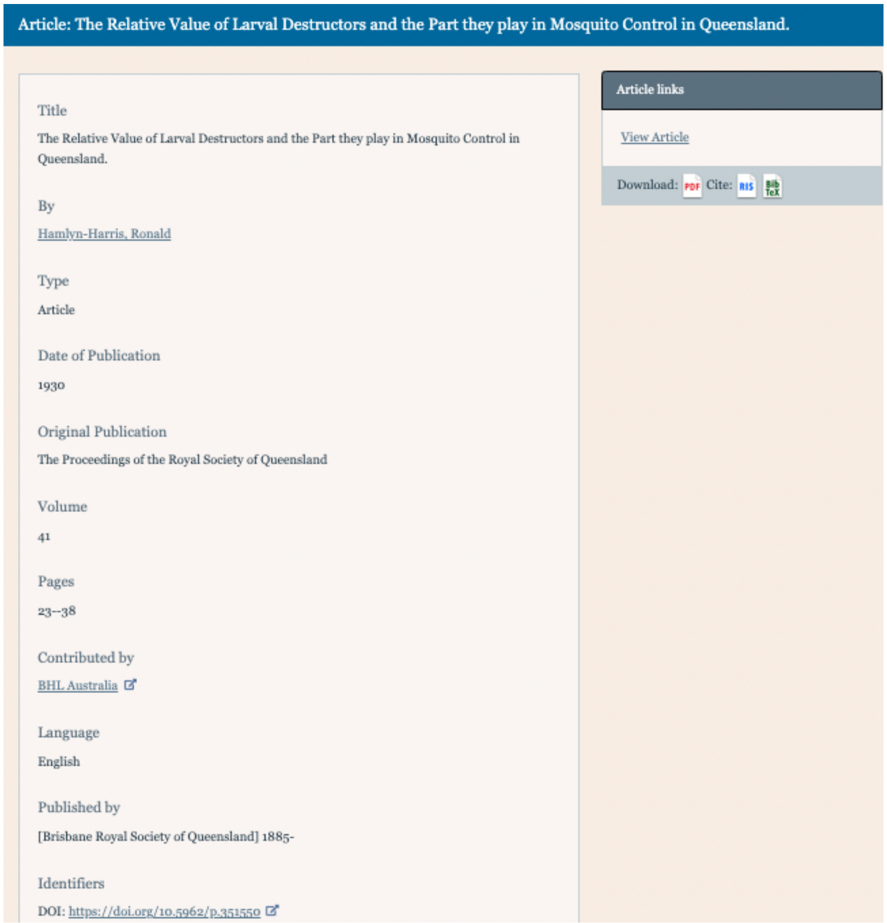
Mentioned by

- 🟥 1 news outlet
- 🟧 1 blog
- 🟦 224 tweeters
- ⬛ 9 Wikipedia pages

Figure: Tracking output and reach. Altmetric brings traditional citation metrics used for scholarly publications together alongside modern analytics from Wikipedia, Wikidata, Wikimedia Commons, and mentions on the web from social media, news, and blogs. (Image: Kearney, 2023)

At a high level, the steps involved in article creation and DOI assignment in BHL are:

1. Generate missing metadata for BHL articles;
2. harvest or upload article metadata as a BHL Part (P6535);
3. register the metadata with Crossref (the DOI registration agency for scholarly content) which then subsequently assigns a DOI to the article deposit;
4. expose a resolvable DOI on the article record (BHL Part)



Article: The Relative Value of Larval Destructors and the Part they play in Mosquito Control in Queensland.

Title

The Relative Value of Larval Destructors and the Part they play in Mosquito Control in Queensland.

By

Hamlyn-Harris, Ronald

Type

Article

Date of Publication

1930

Original Publication

The Proceedings of the Royal Society of Queensland

Volume

41

Pages

23--38

Contributed by

BHL Australia

Language

English

Published by

[Brisbane Royal Society of Queensland] 1885-

Identifiers

DOI: https://doi.org/10.5962/p.351550

Figure: A newly defined article in BHL with its DOI.

This workflow links previously "dark" literature into the modern linked network of scholarly research. Generating the missing metadata (step 1) is unquestionably the most laborious part of the workflow. In 2009, Dr. Rod Page, Professor of Taxonomy at the University of Glasgow and an active member of the PIWG, lamented on his blog that there was a dearth of article metadata in BHL:

> *"BHL has very little article-level metadata, making searching for articles a frustrating experience." (Page, R. D. M., 2009)*

Determined to solve this problem for himself and other users, Dr. Page created BioStor, a web application, that has identified and described over 239,000 articles in BHL. (Page, R. D. M., 2011) Thankfully, BHL quickly partnered with the BioStor project to harvest metadata for articles, bringing that high-value BioStor metadata back into BHL. Dr. Page is an exemplar, showing how vital user feedback is to enhancing BHL's ability to serve the global biodiversity research community.

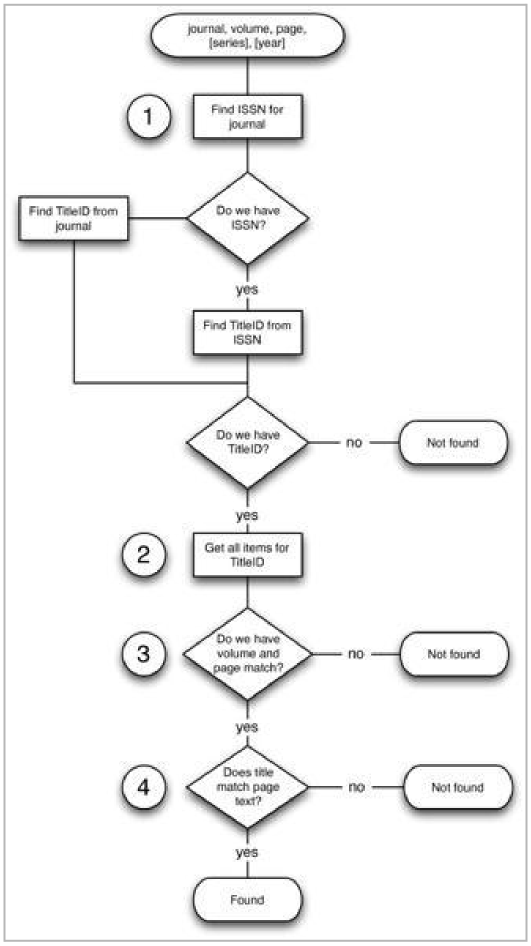

journal, volume, page, [series], [year]
1
Find ISSN for journal
Find TitleID from journal
Do we have ISSN?
yes
Find TitleID from ISSN
Do we have TitleID?
no
Not found
yes
2
Get all items for TitleID
3
Do we have volume and page match?
no
Not found
yes
4
Does title match page text?
no
Not found
yes
Found

Figure: Flow chart of an algorithm for finding articles in BHL. (Image: Page, R. D. M., 2011)

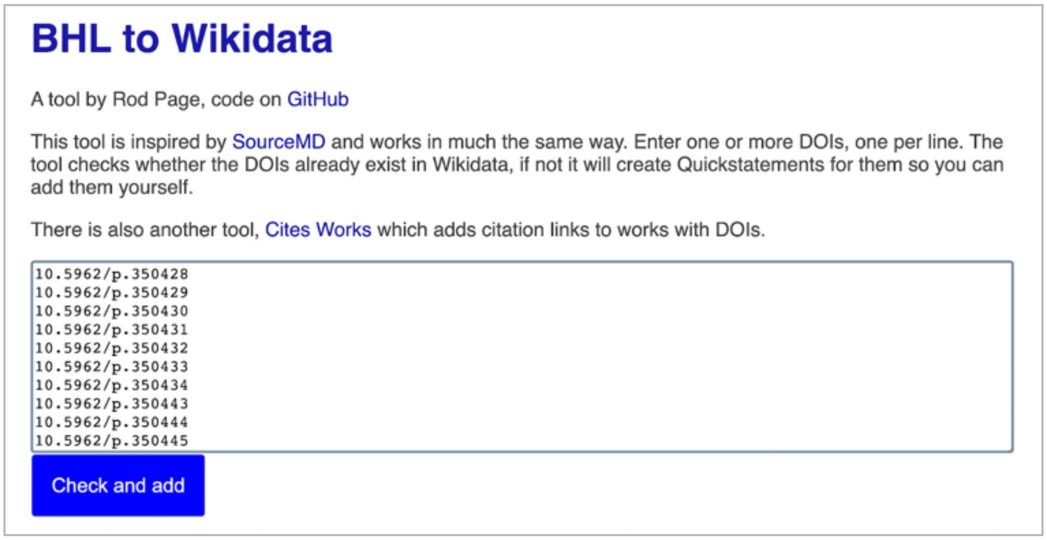
Recently, Dr. Page and members of the PIWG have been testing his new tool called BHL2Wiki to deposit article metadata and their DOIs in Wikidata. While there are some existing bots and tools that bring metadata and DOIs into Wikidata (see: SourceMD), these tools bring author names into Wikidata as “strings,” not “things” (i.e. linked named entities with a URI). The reason the names cannot be reconciled via Wikidata QuickStatements with these tools is that CrossRef only accepts ORCIDs for author names, and ORCIDs are not assigned to a majority of authors in historic literature. Without a persistent identifier in the metadata for authors, no linkage in Wikidata can be made to the author's name.

This means BHL Creator IDs, and all that hard work of BHL’s author disambiguators, is lost. To solve the problem, Rod Page developed the BHL2Wiki tool that, when provided with a list of BHL DOIs, will check that the DOI is not already in Wikidata.

If the DOI is not present in Wikidata, the tool will:

1. pull publication metadata from Crossref,
2. swing past BHL to collect the BHL Creator and Wikidata IDs,
3. produce QuickStatements to be reviewed and batch processed,
4. create new entity records for each publication and its DOI in Wikidata.

The Workflow



BHL to Wikidata

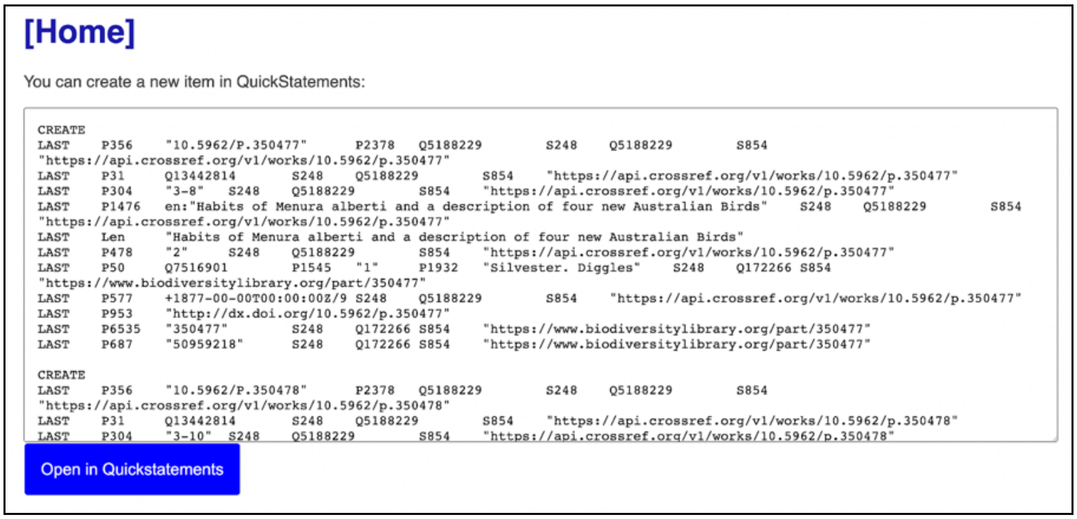
A tool by Rod Page, code on GitHub

This tool is inspired by SourceMD and works in much the same way. Enter one or more DOIs, one per line. The tool checks whether the DOIs already exist in Wikidata, if not it will create Quickstatements for them so you can add them yourself.

There is also another tool, Cites Works which adds citation links to works with DOIs.

```
10.5962/p.350428
10.5962/p.350429
10.5962/p.350430
10.5962/p.350431
10.5962/p.350432
10.5962/p.350433
10.5962/p.350434
10.5962/p.350443
10.5962/p.350444
10.5962/p.350445
```

Figure: Step 1 - Enter a list of DOIs. (Image: Kearney, 2023)



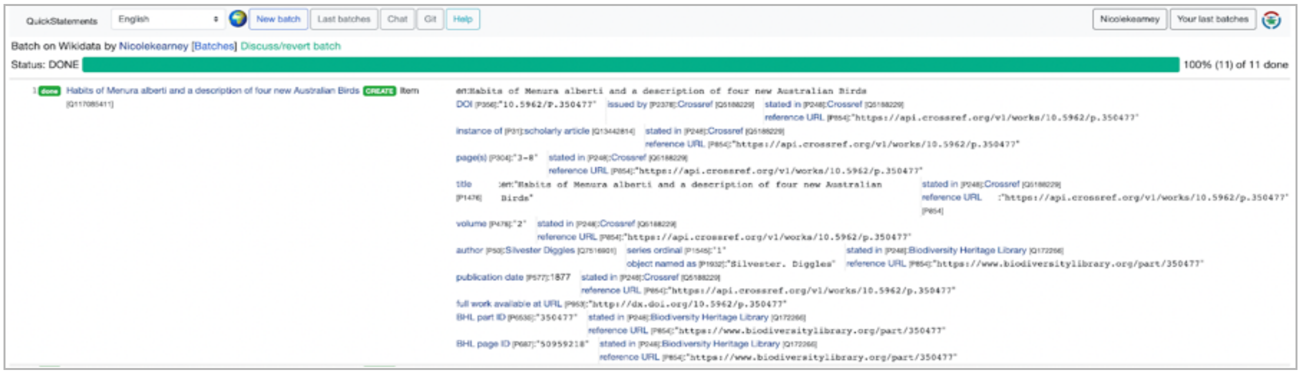
[Home]

You can create a new item in QuickStatements:

```
CREATE
LAST    P356    "10.5962/P.350477"      P2378   Q5188229        S248    Q5188229        S854
"https://api.crossref.org/v1/works/10.5962/p.350477"
LAST    P31     Q13442814       S248    Q5188229        S854    "https://api.crossref.org/v1/works/10.5962/p.350477"
LAST    P304    "3-8"   S248    Q5188229        S854    "https://api.crossref.org/v1/works/10.5962/p.350477"
LAST    P1476   en:"Habits of Menura alberti and a description of four new Australian Birds"    S248    Q5188229        S854
"https://api.crossref.org/v1/works/10.5962/p.350477"
LAST    Len     "Habits of Menura alberti and a description of four new Australian Birds"
LAST    P478    "2"     S248    Q5188229        S854    "https://api.crossref.org/v1/works/10.5962/p.350477"
LAST    P50     Q7516901        P1545   "1"     P1932   "Silvester. Diggles"    S248    Q172266 S854
"https://www.biodiversitylibrary.org/part/350477"
LAST    P577    +1877-00-00T00:00:00Z/9 S248    Q5188229        S854    "https://api.crossref.org/v1/works/10.5962/p.350477"
LAST    P953    "http://dx.doi.org/10.5962/p.350477"
LAST    P6535   "350477"        S248    Q172266 S854    "https://www.biodiversitylibrary.org/part/350477"
LAST    P687    "50959218"      S248    Q172266 S854    "https://www.biodiversitylibrary.org/part/350477"

CREATE
LAST    P356    "10.5962/P.350478"      P2378   Q5188229        S248    Q5188229        S854
"https://api.crossref.org/v1/works/10.5962/p.350478"
LAST    P31     Q13442814       S248    Q5188229        S854    "https://api.crossref.org/v1/works/10.5962/p.350478"
LAST    P304    "3-10"  S248    Q5188229        S854    "https://api.crossref.org/v1/works/10.5962/p.350478"
```

Figure - Step 2: Generate QuickStatements from the DOIs and review Wikidata QuickStatements (Image: Kearney, 2023)



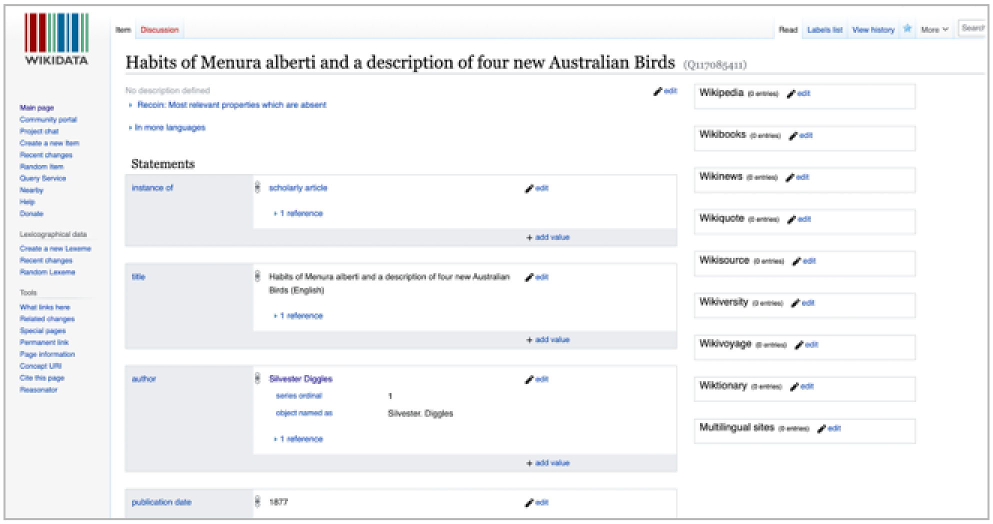
Batch on Wikidata by Nicolekearney [Batches] Discuss/revert batch

Status: DONE 100% (11) of 11 done

1 done Habits of Menura alberti and a description of four new Australian Birds CREATE Item [Q117085411]

en:Habits of Menura alberti and a description of four new Australian Birds
DOI [P356]:"10.5962/P.350477" issued by [P2378]:Crossref [Q5188229] stated in [P248]:Crossref [Q5188229]
reference URL [P854]:"https://api.crossref.org/v1/works/10.5962/p.350477"
instance of [P31]:scholarly article [Q13442814] stated in [P248]:Crossref [Q5188229]
reference URL [P854]:"https://api.crossref.org/v1/works/10.5962/p.350477"
page(s) [P304]:"3-8" stated in [P248]:Crossref [Q5188229]
reference URL [P854]:"https://api.crossref.org/v1/works/10.5962/p.350477"
title [P1476] :en:"Habits of Menura alberti and a description of four new Australian Birds" stated in [P248]:Crossref [Q5188229]
reference URL [P854] :"https://api.crossref.org/v1/works/10.5962/p.350477"
volume [P478]:"2" stated in [P248]:Crossref [Q5188229]
reference URL [P854]:"https://api.crossref.org/v1/works/10.5962/p.350477"
author [P50]:Silvester Diggles [Q7516901] series ordinal [P1545]:"1" stated in [P248]:Biodiversity Heritage Library [Q172266]
object named as [P1932]:"Silvester. Diggles" reference URL [P854]:"https://www.biodiversitylibrary.org/part/350477"
publication date [P577]:1877 stated in [P248]:Crossref [Q5188229]
reference URL [P854]:"https://api.crossref.org/v1/works/10.5962/p.350477"
full work available at URL [P953]:"http://dx.doi.org/10.5962/p.350477"
BHL part ID [P6535]:"350477" stated in [P248]:Biodiversity Heritage Library [Q172266]
reference URL [P854]:"https://www.biodiversitylibrary.org/part/350477"
BHL page ID [P687]:"50959218" stated in [P248]:Biodiversity Heritage Library [Q172266]
reference URL [P854]:"https://www.biodiversitylibrary.org/part/350477"

Figure: Step 3 - Batch process QuickStatements (Image: Kearney, 2023)



Habits of Menura alberti and a description of four new Australian Birds (Q117085411)

No description defined

- Recoin: Most relevant properties which are absent
- In more languages

Statements

instance of: scholarly article — 1 reference

title: Habits of Menura alberti and a description of four new Australian Birds (English) — 1 reference

author: Silvester Diggles
series ordinal: 1
object named as: Silvester. Diggles
— 1 reference

publication date: 1877

Wikipedia (0 entries)
Wikibooks (0 entries)
Wikinews (0 entries)
Wikiquote (0 entries)
Wikisource (0 entries)
Wikiversity (0 entries)
Wikivoyage (0 entries)
Wiktionary (0 entries)
Multilingual sites (0 entries)

Figure: Step 4 - Article and DOI deposits in Wikidata. (Image: Kearney, 2023)

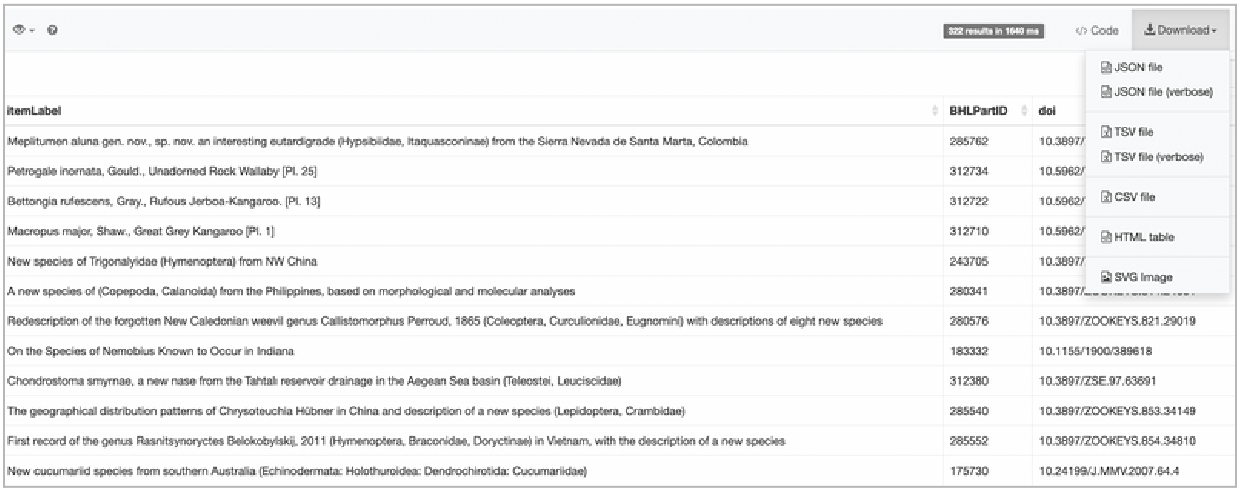
Members of PIWG have astutely recognized that adding BHL metadata and DOIs to Wikidata is yet another important step in their articleization workflows to further increase the accessibility and discoverability of biodiversity literature on the web.

2.2.3 Enabling Interdisciplinary Research with SPARQL

Difficult research questions often require considerable time and labor investments to compile data from disparate sources. The Wikidata Query Service (WDQS) allows users to search across Wikidata's 17.8 billion triples (Haller et al., 2022). The query service was enabled on Wikidata in 2016, providing a user-friendly interface for its SPARQL (pronounced "sparkle") end-point. SPARQL is a query language that allows you to search against semantic databases (W3C, 2013).

Not only does WDQS return raw data with outputs in JSON, TSV, CSV, and HTML but it also has native visualization tools that can render data as a map, image grid, timeline, bubble chart, entity graph, and more. A few examples illustrating the power of the Wikidata Query Service are below:

Raw Data



itemLabel	BHLPartID	doi
Meplitumen aluna gen. nov., sp. nov. an interesting eutardigrade (Hypsibiidae, Itaquasconinae) from the Sierra Nevada de Santa Marta, Colombia	285762	10.3897/
Petrogale inornata, Gould., Unadorned Rock Wallaby [Pl. 25]	312734	10.5962/
Bettongia rufescens, Gray., Rufous Jerboa-Kangaroo. [Pl. 13]	312722	10.5962/
Macropus major, Shaw., Great Grey Kangaroo [Pl. 1]	312710	10.5962/
New species of Trigonalyidae (Hymenoptera) from NW China	243705	10.3897/
A new species of (Copepoda, Calanoida) from the Philippines, based on morphological and molecular analyses	280341	10.3897/
Redescription of the forgotten New Caledonian weevil genus Callistomorphus Perroud, 1865 (Coleoptera, Curculionidae, Eugnomini) with descriptions of eight new species	280576	10.3897/ZOOKEYS.821.29019
On the Species of Nemobius Known to Occur in Indiana	183332	10.1155/1900/389618
Chondrostoma smyrnae, a new nase from the Tahtalı reservoir drainage in the Aegean Sea basin (Teleostei, Leuciscidae)	312380	10.3897/ZSE.97.63691
The geographical distribution patterns of Chrysoteuchia Hübner in China and description of a new species (Lepidoptera, Crambidae)	285540	10.3897/ZOOKEYS.853.34149
First record of the genus Rasnitsynoryctes Belokobylskij, 2011 (Hymenoptera, Braconidae, Doryctinae) in Vietnam, with the description of a new species	285552	10.3897/ZOOKEYS.854.34810
New cucumariid species from southern Australia (Echinodermata: Holothuroidea: Dendrochirotida: Cucumariidae)	175730	10.24199/J.MMV.2007.64.4

Figure: Raw Data Exports of BHL Parts and Pages with DOIs in Wikidata (Image: Dearborn, 2022 https://w.wiki/6Y$B)

Maps

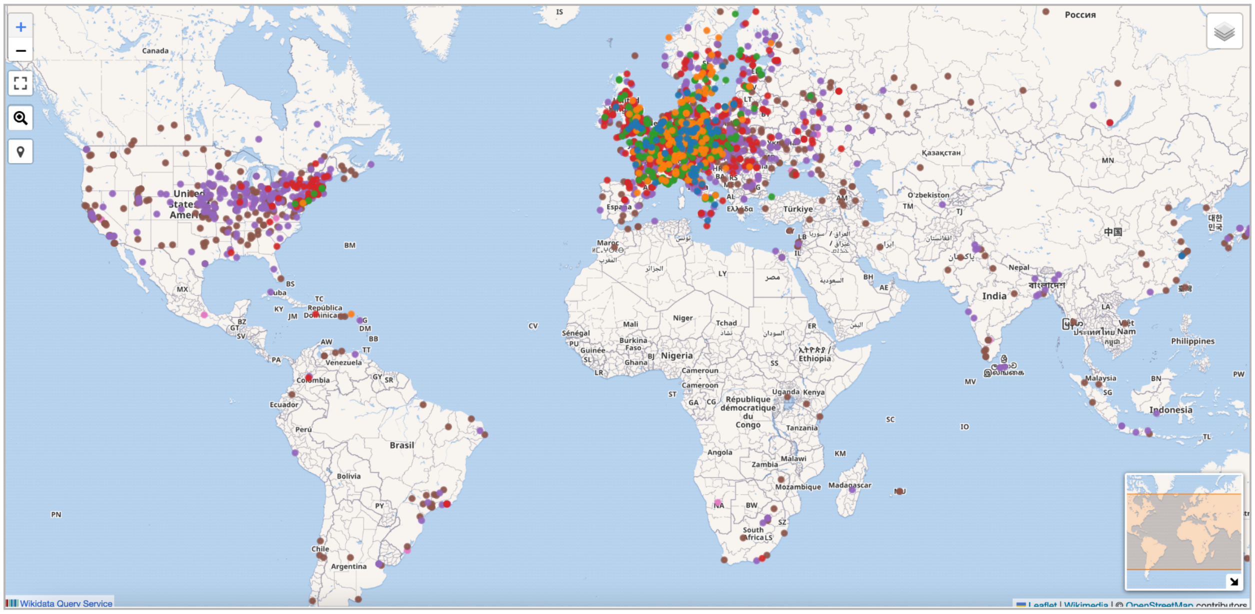


Figure: A map of deceased entomologists' birthplaces, color-coded by era. (Image: Dearborn, 2022 https://w.wiki/4JFr)

Timelines

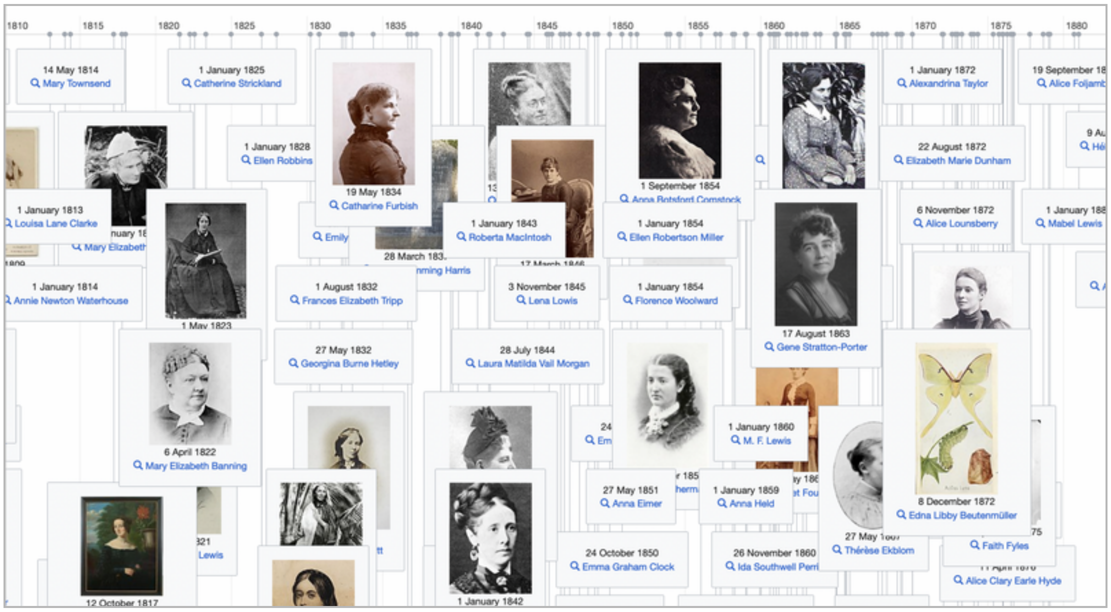


Figure: Birthdays of female scientific illustrators in BHL (Image: Dearborn, 2022 https://w.wiki/4dwD)

Entity Graphs

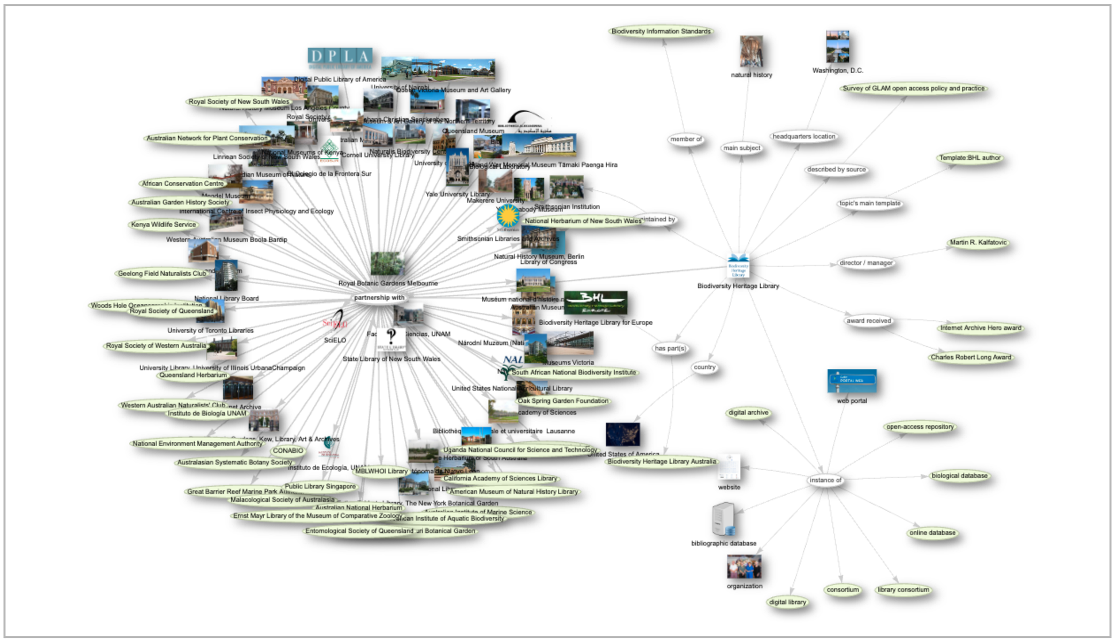

Biodiversity Information Standards
natural history
Washington, D.C.
Survey of GLAM open access policy and practice
DPLA
Digital Public Library of America
Royal Society of New South Wales
Australian Network for Plant Conservation
Queensland Museum
member of
main subject
headquarters location
Template:BHL author
described by source
topic's main template
African Conservation Centre
Australian Garden History Society
Kenya Wildlife Service
Auckland War Memorial Museum Tāmaki Paenga Hira
Yale University Library
Smithsonian Institution
National Herbarium of New South Wales
maintained by
Smithsonian Libraries
Martin R. Kalfatovic
Geelong Field Naturalists Club
Royal Botanic Gardens Melbourne
partnership with
Natural History Museum, Berlin
Library of Congress
Biodiversity Heritage Library
director / manager
Royal Society of Queensland
University of Toronto Libraries
SciELO
Biodiversity Heritage Library for Europe
award received
Internet Archive Hero award
Royal Society of Western Australia
has part(s)
Charles Robert Long Award
University of Illinois UrbanaChampaign
Queensland Herbarium
State Library of New South Wales
Museums Victoria
South African National Biodiversity Institute
country
Western Australian Naturalists' Club
Instituto de Biología UNAM
United States National Agricultural Library
Oak Spring Garden Foundation
web portal
digital archive
open-access repository
National Environment Management Authority
CONABIO
Australasian Systematic Botany Society
Uganda National Council for Science and Technology
United States of America
Biodiversity Heritage Library Australia
website
instance of
biological database
Great Barrier Reef Marine Park Authority
Public Library Singapore
California Academy of Sciences Library
American Museum of Natural History Library
Malacological Society of Australasia
Australian National Herbarium
The New York Botanical Garden
Ernst Mayr Library of the Museum of Comparative Zoology
Entomological Society of Queensland
online database
bibliographic database
organization
digital library
consortium
library consortium

Figure: An entity graph of the BHL's Partner Network; visualizing Property:P2652 "partnership with." (Image: Dearborn, 2022 https://w.wiki/4G3B)

What other questions might be asked of BHL's data with the potential to yield new insights about the natural world?

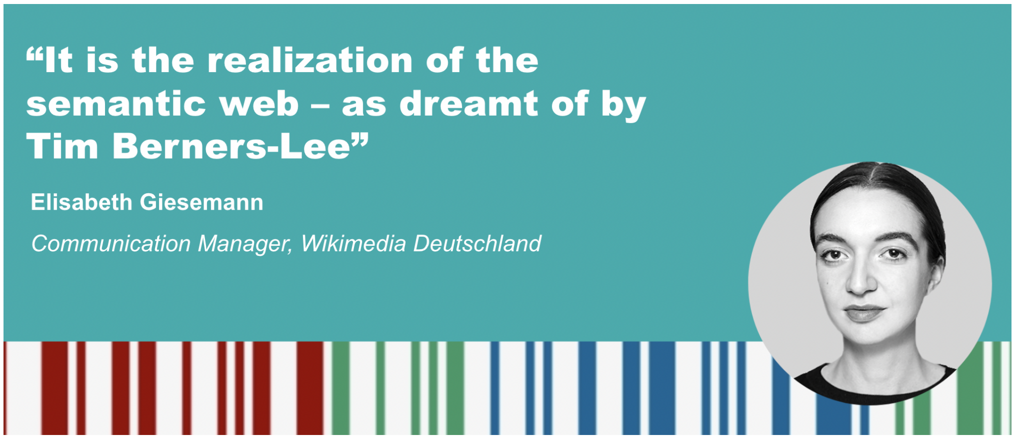
The Wikidata Query Service could also be leveraged for complex collections analysis and help to identify gaps in content coverage which could inform collection development and digitization activities. Sample questions might include:

- *Which publications in BHL have the greatest frequency of species descriptions and /or nomenclatural acts?*
- *May I have a list of works published in / or about specific biodiversity hotspots e.g. the African Great Lakes region? The Congo Basin? Upper Amazonia? New Guinea?*
- *How many authors in BHL are women? – Women entomologists; botanists; herpetologists? collectors; scientific illustrators?*

With WDQS, data from a variety of knowledge domains can inform interdisciplinary research questions. Subsequent data submission and linking work can advance opportunities for complex content analysis of the BHL corpus.

2.3 Future Outlook

As the world's largest openly editable knowledge base, Wikidata, and its underlying software Wikibase, present a compelling vision promising to de-silo all of the world's data.



"It is the realization of the semantic web – as dreamt of by Tim Berners-Lee"

Elisabeth Giesemann

Communication Manager, Wikimedia Deutschland

Figure: In a recent talk given to data journalists, Elisabeth Giesemann, Communication Manager for Wikimedia Deutschland, describes Wikidata as "the realization of the semantic web – as dreamt of by Tim Berners-Lee." (Chaos Computer Club, 2021)

Nevertheless, achieving this dream does not come without challenges:

> *"The project suffers from the biases and vandalism that plague other Wikimedia projects. Including gender gaps in the contributor base—the majority of the volunteer editors are male. And the majority of the data is from—and about—the Northern Hemisphere. The project is young, Giesemann emphasizes" (Sengel-Jones, 2021).*

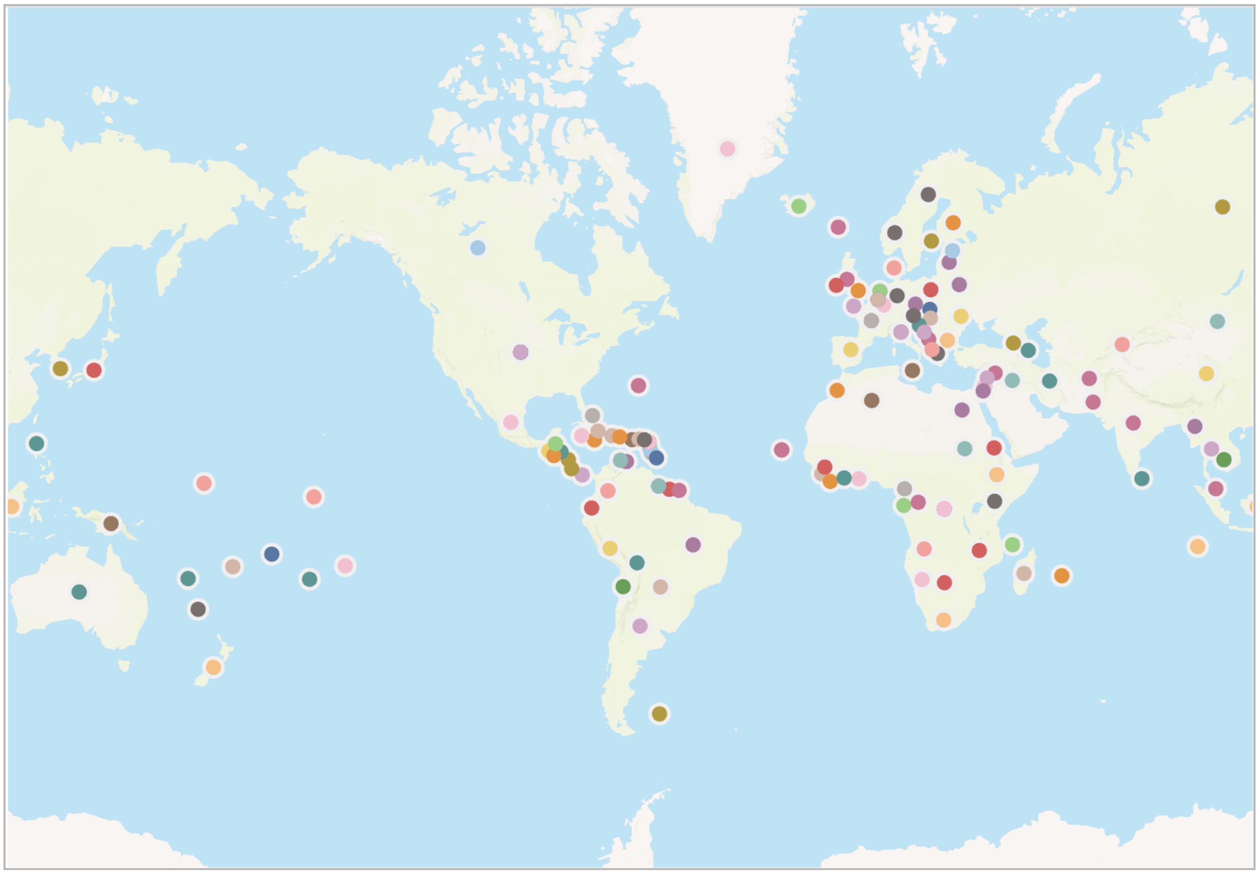
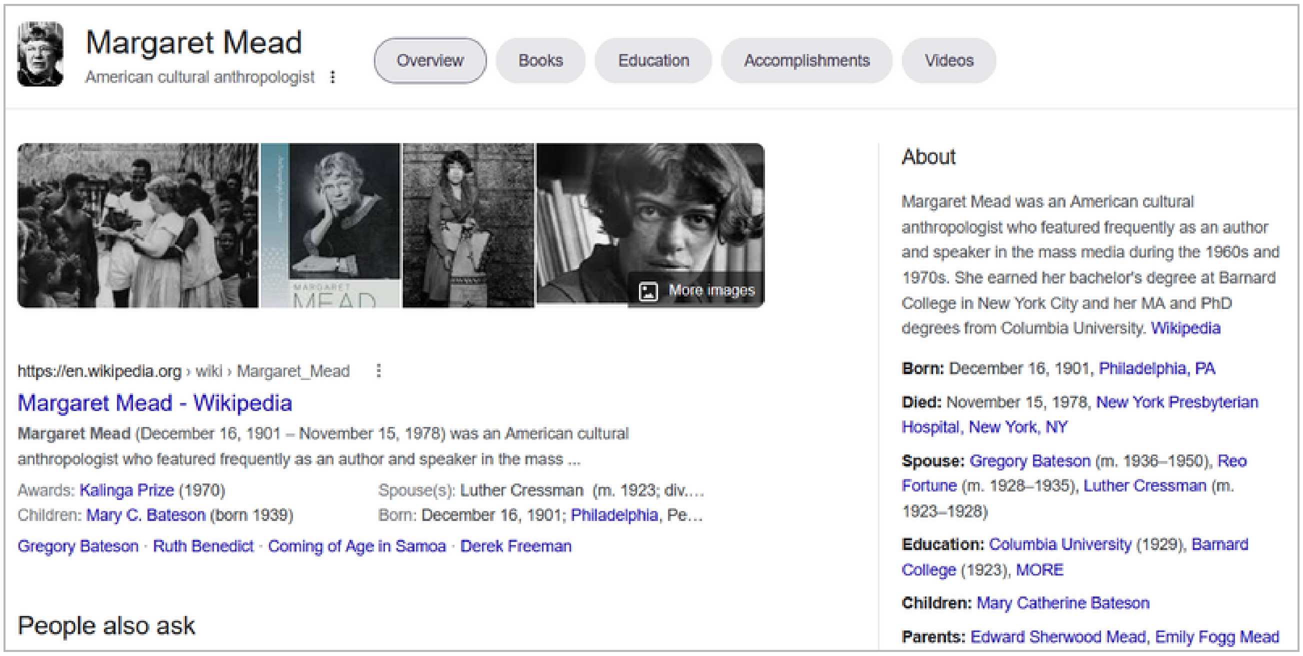


Figure: A BHL topical coverage map shows many publications in BHL that could help fill knowledge representation gaps by focusing on vulnerable biodiversity hotspots in Africa and Oceania. (World's Most Vulnerable Nations Suffer Disproportionately, n.d.)(Image: Dearborn 2022)

Despite its detractors, Wikidata holds extraordinary promise to break down data silos. This growing global data hub is now an intimate part of our daily lives. Alexa, Siri, and Google all use Wikidata to drive their ubiquitously relied-upon search services. Google decommissioned its knowledge graph database, Freebase, and in 2014 opted to migrate the data to Wikidata (Pellissier Tanon et al., 2016). Today, Google Search relies heavily on data from Wikidata, Wikipedia, schema.org microdata, and other licensed data sources to drive its knowledge graph panels. (Google, 2023)



Margaret Mead
American cultural anthropologist

https://en.wikipedia.org › wiki › Margaret_Mead

Margaret Mead - Wikipedia

Margaret Mead (December 16, 1901 – November 15, 1978) was an American cultural anthropologist who featured frequently as an author and speaker in the mass ...

Awards: Kalinga Prize (1970)
Children: Mary C. Bateson (born 1939)
Spouse(s): Luther Cressman (m. 1923; div.…
Born: December 16, 1901; Philadelphia, Pe…

Gregory Bateson · Ruth Benedict · Coming of Age in Samoa · Derek Freeman

People also ask

About

Margaret Mead was an American cultural anthropologist who featured frequently as an author and speaker in the mass media during the 1960s and 1970s. She earned her bachelor's degree at Barnard College in New York City and her MA and PhD degrees from Columbia University. Wikipedia

Born: December 16, 1901, Philadelphia, PA

Died: November 15, 1978, New York Presbyterian Hospital, New York, NY

Spouse: Gregory Bateson (m. 1936–1950), Reo Fortune (m. 1928–1935), Luther Cressman (m. 1923–1928)

Education: Columbia University (1929), Barnard College (1923), MORE

Children: Mary Catherine Bateson

Parents: Edward Sherwood Mead, Emily Fogg Mead

Figure: A Google Knowledge Graph Panel of Margaret Mead; much of the data that populates the panel is sourced from Wikidata. (Image: Google, n.d.)

As a central data repository that drives core third-party Internet services, the quality, coverage, and structuring of data in Wikidata have never mattered more. At its best Wikidata is opening new knowledge pathways, connecting disparate ideas, and helping uncover untold stories from the margins. At its worst, Wikidata can reinforce old power structures, highlight gaps in human knowledge, and reduce complexity to the point of harm (Keyes, 2019).

BHL as a data source holds immense untapped potential and would expand Wikidata's quality and breadth in the life science domain. Additionally, according to Andy Mabbett, Independent Wikimedian, the Wikispecies project is another rich source of biodiversity data and persistent identifiers that should be migrated to Wikidata for the multilingual, searching, and interlinking advantages. BHL staff have a deep bench of advanced data modeling and crosswalking expertise and are ready to assist with biodiversity data enrichment projects and ontology refinements.

Wikidata may be humanity's opportunity to re-imagine and re-frame knowledge representation — and in so doing, perhaps the chance to change our collective narrative to respect, include, and celebrate the diversity of all life on Earth.

> *"Until Wikidata can give me a list of all movies shot in the 60s in Spain, in which a black female horse is stolen by a left-handed actor playing a Portuguese orphan, directed by a colorblind German who liked sailing, and written by a dog-owning women [sic] from Helsinki, we have more work to do." (User: Tobias1984, Wikidata project chat, 2013)*

3.0 Wikibase

3.1 About

Like all Wikimedia projects, Wikidata is built on MediaWiki, an open-source wiki software with over 1,500 extensions. (Wikimedia Foundation, 2022) Wikibase is just one of many MediaWiki extensions (actually a suite of extensions) written primarily in PHP and developed for the Wikidata project, providing collaborative editing and storage for structured linked data. Wikibase supports custom ontologies, SPARQL queries, federated queries (searching across many Wikibase instances), and data exports in multiple file formats: XML, RDF, and JSON (Wikidata, 2017).

Alongside the Wikidata community, the Wikibase development community continues to grow and enhance the open-source software's offerings. The Wikimedia Deutschland Team hopes to make Wikibase installations more accessible to non-technical audiences by building new community tools and offering cloud-hosted solutions. In the recent publication "Strategy for the Wikibase Ecosystem," the vision for the platform is bold — placing Wikibase at the heart of the rapidly unfolding decentralized semantic web, making data exchange seamless across a universe of Wikibase nodes. The Wikibase Strategy opens with this quote:

> *"The internet explodes when somebody has the creativity to look at a piece of data that's put there for one reason and realise they can connect it with something else." –Sir Tim Berners-Lee*
> *(Pintscher et al., 2019)*

Currently, there are three install methods and a fourth, invite-only hosted solution: Wikibase.cloud (Wikibase, 2023).

Some GLAMs are opting for a separate Wikibase instance to take advantage of:

- Control over data quality,
- granular user permissions,
- data privacy for in-copyright data or personally identifiable information (PII),
- custom data modeling; use of domain-specific ontologies, and
- potential for SPARQL query federation with Wikidata and other Wikibase instances.

Naturally, the promise of search federation across a global network of Wikibase instances is exciting but note that federation options, like many Wikibase features, are still in very early stages of development and unless you are using the Wikidata ontology or appear on this list of federated knowledge base endpoints, your options for data interoperability are limited. (Wikimedia Foundation, 2022). Through query federation, the Wikimedia Deutschland Team envisions a global linked-open data ecosystem being erected using Wikibase.

> *"...we imagine that one day all the Wikibase instances will be connected between themselves and back to Wikidata. "*
>
> *– Wikimedia Deutschland | Tech News*

A current focus for Wikimedia developers is to make Wikibase deployments more accessible to organizations that lack resources and technical capacity. Additionally, the whole community is dealing with scaling issues for the Wikidata Query Service that relies on Blazegraph, a triple-store database that has reached end-of-life. Not only is Blazegraph at capacity and suffering from performance issues, but its codebase also went dormant in 2018 (after its acquisition by Amazon). The Wikimedia Search Team is seeking a suitable replacement for Blazegraph. The top replacement candidates are Jena and Virtuoso but data migration has yet to be completed. (WDQS Search Team, 2022) Migrating off of Blazegraph is crucial for the project's overall sustainability.

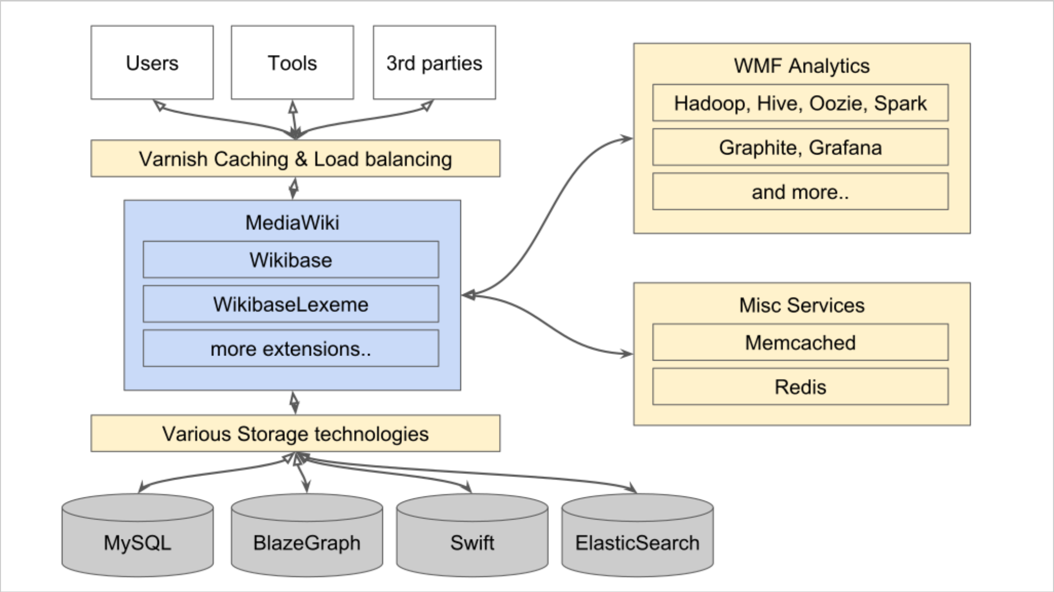


Figure: High-level Wikibase architecture overview (Image: Addshore, 2022)

Despite architecture growing pains, many intrepid GLAMs are experimenting with Wikibase.cloud and/or the Wikibase Docker install option.

3.2 Identified Use Cases

3.2.1 Extending BHL's Data Model

A Wikibase deployment is another pathway for BHL's semantic enrichment. Beyond the current BHL work in Wikidata, Wikibase presents a flexible platform for entity description, data modeling, and reconciliation. Other entities (i.e. "things" that we want to surface and describe in BHL) that are currently under-described in BHL could be modeled by librarians and information architects, enhanced by trained staff and volunteers, and subsequently federated with the rest of the Wikidata ecosystem (with the caveat that federation features develop to accommodate and map to external ontologies).

There are long-standing BHL user requests to collect more robust metadata for and provide search access points to:

- Species descriptions (sometimes referred to as taxonomic treatments)
- Visual media (illustrations, photos, art, maps)

- Geographic locations
- Visual artists (illustrators, lithographers, engravers, painters)
- Specimen field collectors
- Publishers

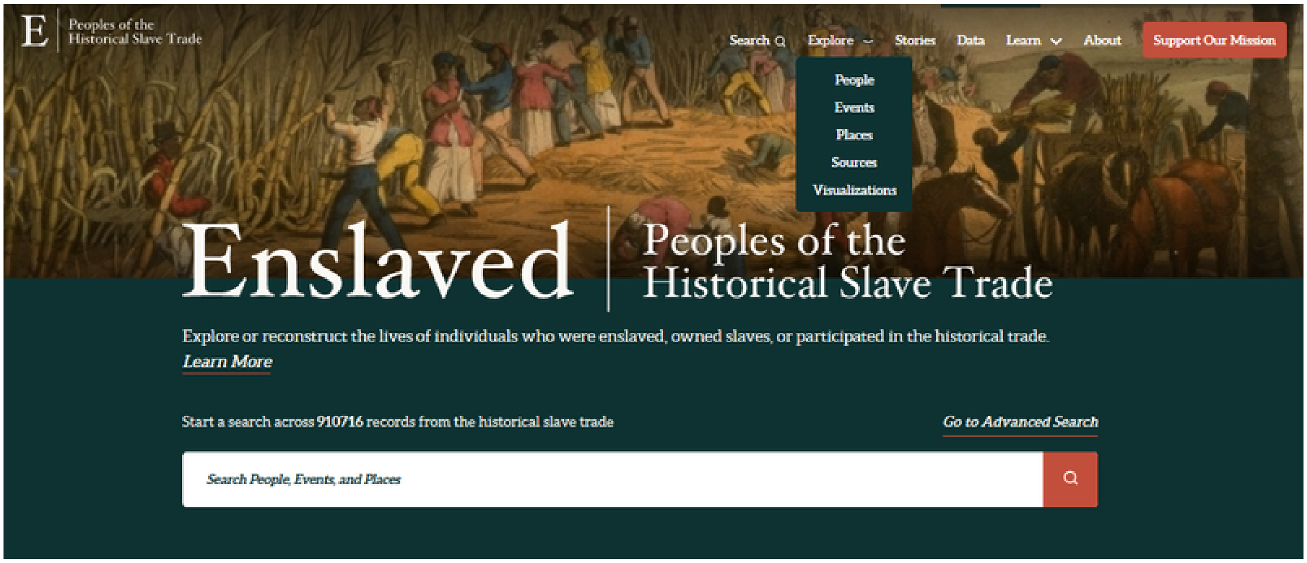
The description of these undescribed entities could happen in Wikibase, without requiring any changes to BHL's current information architecture. An external Wikibase provides a flexible space to experiment and construct data models on the fly without the need for major code changes to the BHL platform itself.

In turn, a Wikibase repository could also provide BHL developers with improved data through staff curation efforts that could be harvested back into the BHL production environment, if desired. BHL staff would be empowered to take a more active role in co-creating a more robust, representative data model for BHL with the eventual aim of releasing the curation of that data to a larger set of enthusiastic volunteers who want nothing more than to see all of BHL's content made useful, usable, and used.

3.2.3 Wikibase Front-end Showcase

Additionally, many projects are using Wikibase as a back-end with slick front-end user interfaces, driven by SPARQL's powerful search features.

Enslaved.org



Enslaved | Peoples of the Historical Slave Trade

Explore or reconstruct the lives of individuals who were enslaved, owned slaves, or participated in the historical trade.

Learn More

Start a search across 910716 records from the historical slave trade

Go to Advanced Search

Search People, Events, and Places

Figure: Enslaved.org developed a custom ontology to model previously underrepresented data. See: "Stories of the Enslaved told using Wikibase."

Linked Jazz

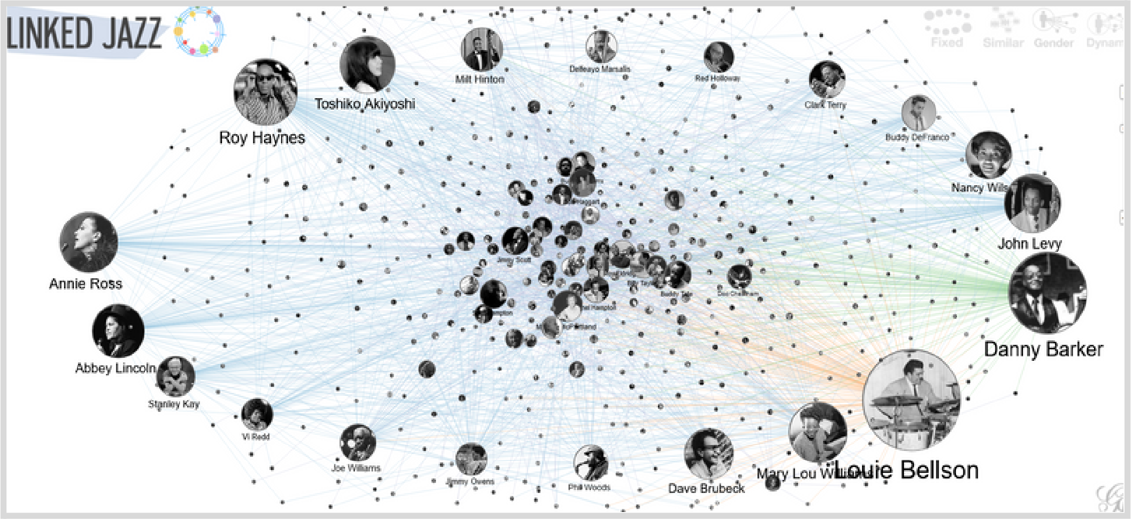
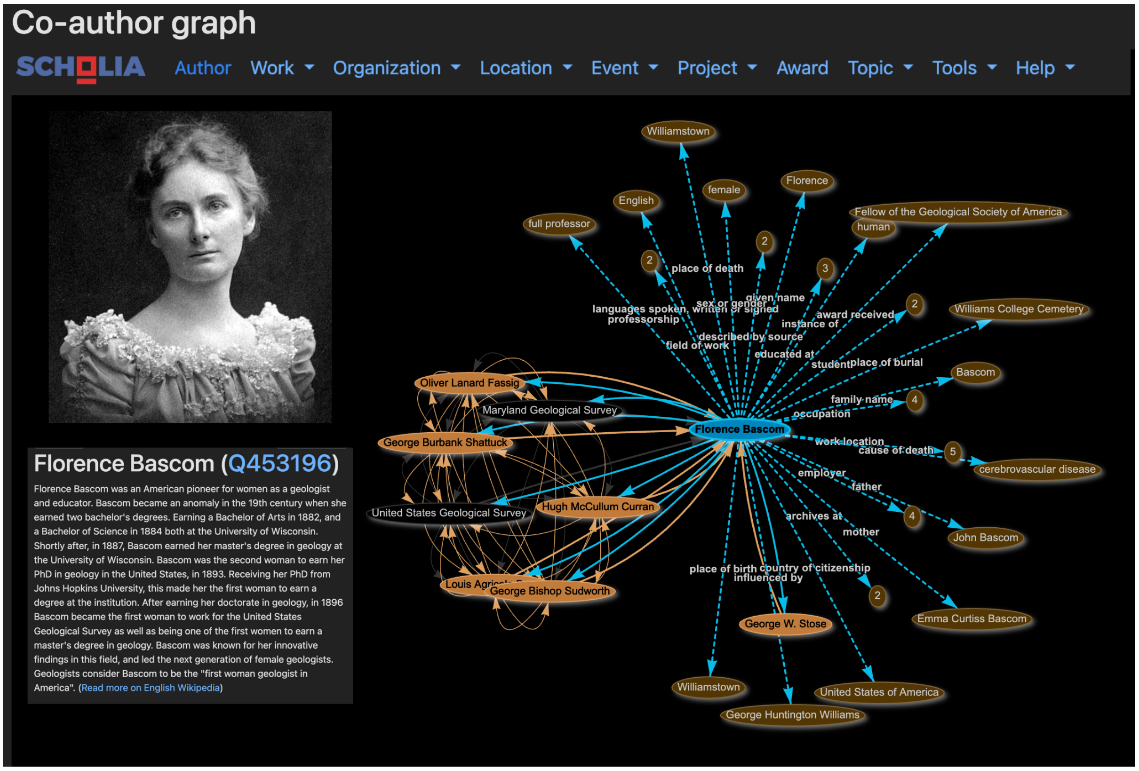


Figure: Pratt's [Linked Jazz](#) project crowdsources relationships between musicians, exposing their community as a knowledge graph user analysis of jazz history materials. (Pratt University, Semantic Lab, n.d.)

Scholia



Florence Bascom (Q453196)

Florence Bascom was an American pioneer for women as a geologist and educator. Bascom became an anomaly in the 19th century when she earned two bachelor's degrees. Earning a Bachelor of Arts in 1882, and a Bachelor of Science in 1884 both at the University of Wisconsin. Shortly after, in 1887, Bascom earned her master's degree in geology at the University of Wisconsin. Bascom was the second woman to earn her PhD in geology in the United States, in 1893. Receiving her PhD from Johns Hopkins University, this made her the first woman to earn a degree at the institution. After earning her doctorate in geology, in 1896 Bascom became the first woman to work for the United States Geological Survey as well as being one of the first women to earn a master's degree in geology. Bascom was known for her innovative findings in this field, and led the next generation of female geologists. Geologists consider Bascom to be the "first woman geologist in America". (Read more on English Wikipedia)

Figure: [Florence Bascom's co-author graph on Scholia.](#) Scholia uses the Wikidata Query Service to build dynamic pages consisting of lists and *visualizations, for researchers, organizations, journals, publishers, scholarly works, and research topics including taxons.*

3.3. Future outlook

The promise of Wikibase has caught the attention of many national libraries including the German National Library, among the first to use Wikibase to host an integrated authority file. Other European libraries including France, Spain, Italy, Wales, Sweden, Luxembourg, and the Netherlands have followed suit with their pilots. (Pintscher et al., 2019)

Nevertheless, the development of Wikibase is still in its nascent stages and for many libraries that lack a dedicated staff of developers, UX designers, information architects, and technical project managers, deployment and adoption is a very steep climb. Moreover, the promise of data interoperability across multiple Wikibases still needs further development and should include an accessible front-end interface for lay users. Accommodating other ontologies should also be considered, in particular BIBFRAME and CIDOC Conceptual Reference Model (CRM), two semantic data models from the cultural heritage sector.

The good news is that BHL joined the Wikibase Stakeholder Group (WBSG) in 2022 to join a group of like-minded organizations banning together to scope and test out new Wikibase features that support community-driven use cases. WBSG organizations aim to pool resources, to serve their collective users' needs. Their work includes many investigations around Wikibase federated search: how it will work, and what functionality will be needed to ensure that Wikibase instances that model their data in a diversity of ways can still "talk" to each other.

In short, Wikibase is certainly worth watching. The promise of the community's vision to unify global knowledge through a distributed network of connected Wikibase instances is so compelling, it should not be ignored.

> *If you expose your data for other people to consume then other people can improve that data for you. New knowledge can then be synthesized out of existing knowledge. [...] Putting your data in a format that other people recognize and expect (like semantic-linked open data) makes it actionable and usable. – James Hare, Internet Archive Wikibase Developer. (Appendix 2: Interviews)*

4.0 Wikimedia Commons

4.1 About

Alongside Wikidata, Wikimedia Commons is another part of the backbone infrastructure that drives content creation on Wikipedia and other Wikimedia sister projects. (Wikimedia Contributors, 2023)

Wikimedia Commons is an open access media repository providing visual source material for all Wikimedia projects. While Wikimedia Commons hosts audio, video, animation, digital books, 3D structure files, and even tabular files — a vast majority of the media found on Commons are image files; 94.5% of files are SVG, PNG, and JPEG formats.

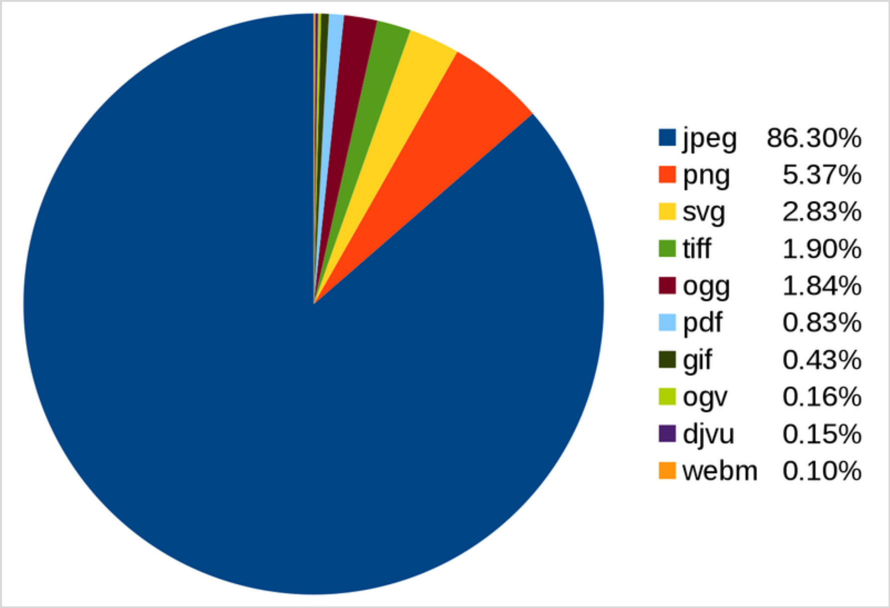
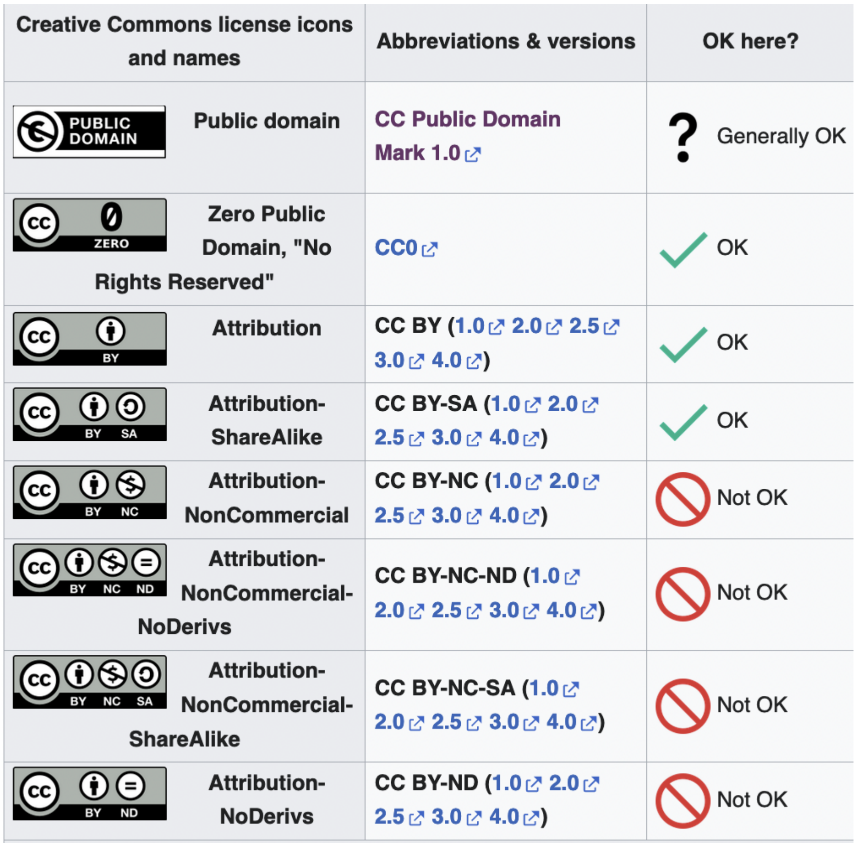


Figure: Percentages of file types on Commons (Image:Delphi234 Vector: MennasDosbin)

To use a media file on any Wikimedia project, it *must* be uploaded to the media repository first under an open license (Commons: Licensing, n.d.).



Creative Commons license icons and names	Abbreviations & versions	OK here?
Public domain	CC Public Domain Mark 1.0	? Generally OK
Zero Public Domain, "No Rights Reserved"	CC0	✓ OK
Attribution	CC BY (1.0 2.0 2.5 3.0 4.0)	✓ OK
Attribution-ShareAlike	CC BY-SA (1.0 2.0 2.5 3.0 4.0)	✓ OK
Attribution-NonCommercial	CC BY-NC (1.0 2.0 2.5 3.0 4.0)	⊘ Not OK
Attribution-NonCommercial-NoDerivs	CC BY-NC-ND (1.0 2.0 2.5 3.0 4.0)	⊘ Not OK
Attribution-NonCommercial-ShareAlike	CC BY-NC-SA (1.0 2.0 2.5 3.0 4.0)	⊘ Not OK
Attribution-NoDerivs	CC BY-ND (1.0 2.0 2.5 3.0 4.0)	⊘ Not OK

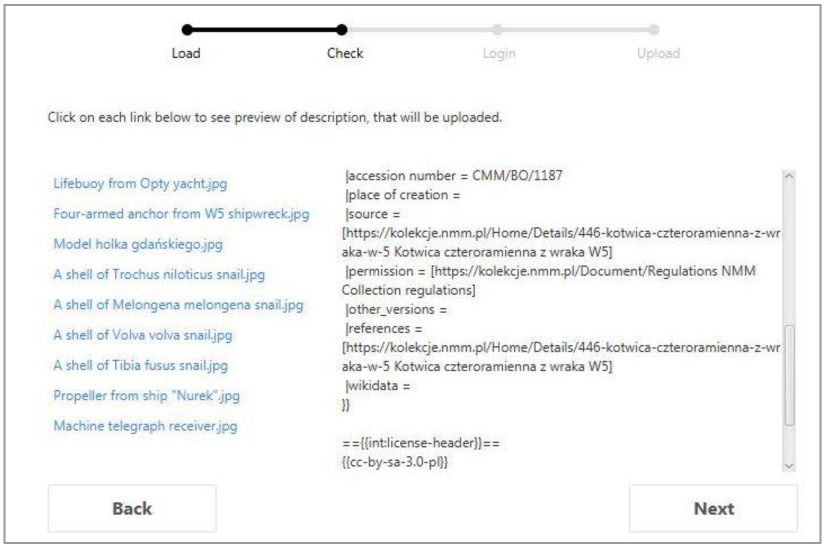
Figure: Wikimedia Commons license requirements. (Image: Commons: Licensing, n.d.)

In addition to the licensing requirements, files must:

- be a media file
- be of an allowable free file format
- be realistically useful for an educational purpose

(Wikimedia Commons, 2022)

Uploading images to Wikimedia Commons is a relatively straightforward process and a multitude of image upload tools are available. There is the standard upload wizard for single image uploads; for bulk uploads of image sets, PattyPan is a widely used tool (Pattypan, 2016). Others have opted for sophisticated scripts to load images to the repository from sources like Flickr that periodically check for new images as they become available.



Click on each link below to see preview of description, that will be uploaded.

- Lifebuoy from Opty yacht.jpg
- Four-armed anchor from W5 shipwreck.jpg
- Model holka gdańskiego.jpg
- A shell of Trochus niloticus snail.jpg
- A shell of Melongena melongena snail.jpg
- A shell of Volva volva snail.jpg
- A shell of Tibia fusus snail.jpg
- Propeller from ship "Nurek".jpg
- Machine telegraph receiver.jpg

```
 |accession number = CMM/BO/1187
 |place of creation =
 |source =
[https://kolekcje.nmm.pl/Home/Details/446-kotwica-czteroramienna-z-wr
aka-w-5 Kotwica czteroramienna z wraka W5]
 |permission = [https://kolekcje.nmm.pl/Document/Regulations NMM
Collection regulations]
 |other_versions =
 |references =
[https://kolekcje.nmm.pl/Home/Details/446-kotwica-czteroramienna-z-wr
aka-w-5 Kotwica czteroramienna z wraka W5]
 |wikidata =
}}

=={{int:license-header}}==
{{cc-by-sa-3.0-pl}}
```

Figure: Pattypan - verifying upload for another batch of files (Image: Marta Malina Moraczewska)

At file upload, users add metadata and assign media to categories. Uploaders may reuse existing categories or create their own. Categories are a product of community-driven folksonomy which, while whimsical and empowering, result in reduced search precision and retrieval in comparison to the application of carefully curated controlled vocabularies. Additionally, the lack of metadata requirements at the point of upload has impacted data quality across the repository.

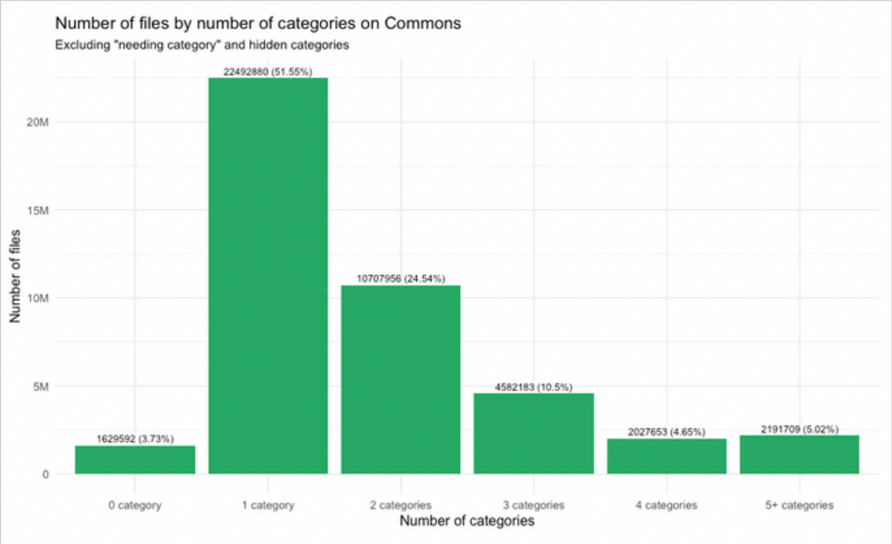


Image by Chelsy Xie/Wikimedia Foundation, CC BY-SA 4.0.

Figure: Categories act like library subject headings, but they are not *required fields. Therefore, most images have only one assigned category. AI and machine learning should be tactics employed to remedy the sparse metadata. (Image: Vershbow & Fauconnier, 2019)*

Metadata quality is a long-standing known issue for the Commons community. Additionally, the inability to translate the metadata into other language editions has been an impediment to image reuse across Wikipedia language editions. To address these challenges, the Wikimedia Foundation received USD 3,015,000 from the Alfred P. Sloan Foundation in 2016 to accelerate the development of Structured Data on Commons in a three-year time frame. (Sloan Grant, 2016) The goal is to provide the Commons community with a way to add machine-readable, structured data about the media files using Wikibase and Wikidata as the underlying data repositories, thereby dramatically improving search functionality across the repository.

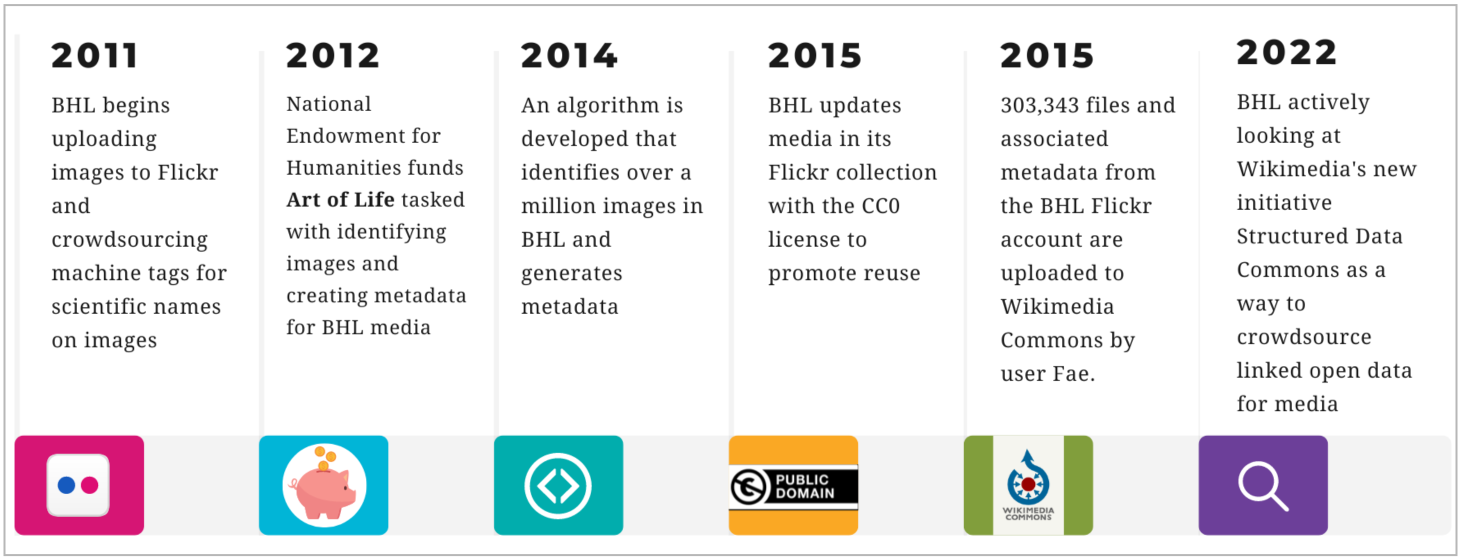
4.2 Identified Use Cases

4.2.1 Image Search with Structured Data Commons (SDC)

BHL's user community has had a long-standing desire for native image search. Millions of scientific illustrations, drawings, paintings, diagrams, maps, charts, tables, graphs, posters, photographs, and other biodiversity visual media exist within the BHL corpus. But since project inception, the metadata needed to drive image search has been missing.

In response, BHL has made great strides in collecting metadata for its images through various initiatives, which include BHL Flickr, a grant from the National Endowment for Humanities called "Art of Life," and the eventual bulk upload to Flickr and Wikimedia Commons. Over the years, these efforts have gone a long way to unlock and expose BHL's media to a larger digital audience.

With the recent departure of Communications Manager, Grace Costantino, who spearheaded BHL's media initiatives, the time has come to take stock of past efforts. (Appendix 2: Interviews) BHL should be exploring new ways to bring images and metadata into a central place and begin the work of normalizing, enhancing, and converting the media asset metadata into linked open data for greater access and dissemination on the web.



2011
BHL begins uploading images to Flickr and crowdsourcing machine tags for scientific names on images

2012
National Endowment for Humanities funds **Art of Life** tasked with identifying images and creating metadata for BHL media

2014
An algorithm is developed that identifies over a million images in BHL and generates metadata

2015
BHL updates media in its Flickr collection with the CC0 license to promote reuse

2015
303,343 files and associated metadata from the BHL Flickr account are uploaded to Wikimedia Commons by user Fae.

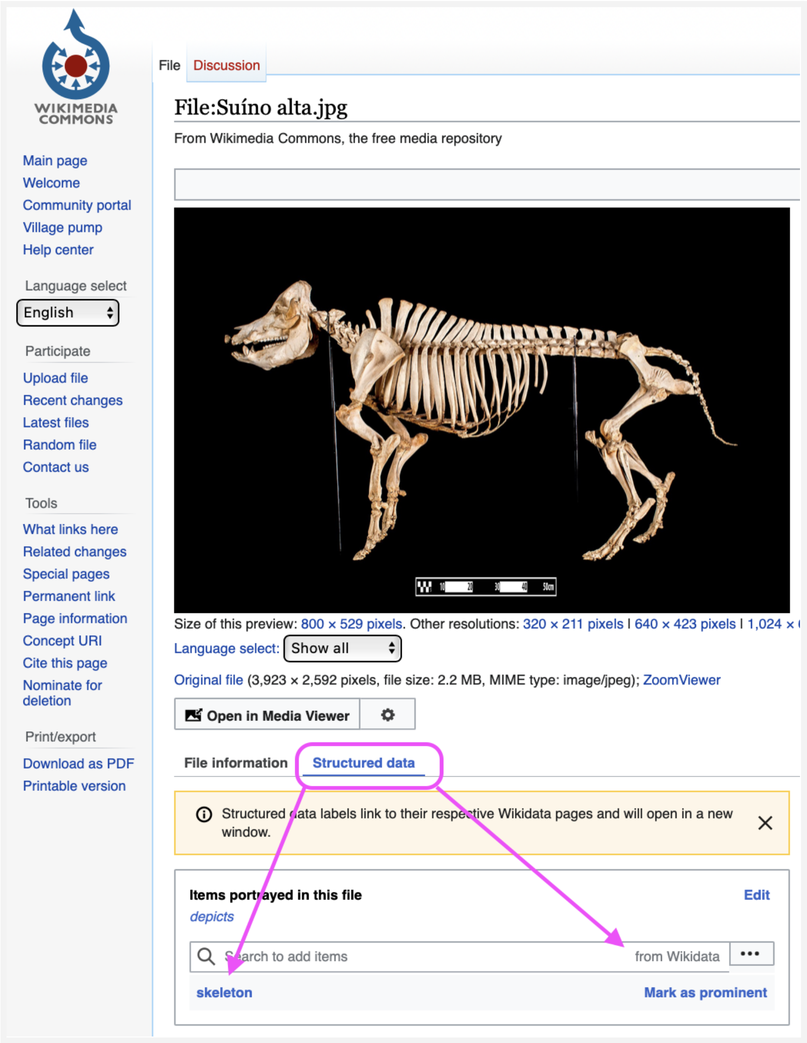
2022
BHL actively looking at Wikimedia's new initiative Structured Data Commons as a way to crowdsource linked open data for media

Figure: Timeline of BHL image metadata capture. (Image: Dearborn & Kalfatovic, 2022)

Structured Data Commons (SDC), recently launched on Wikimedia Commons, presents the perfect place for image metadata enhancement to continue. SDC is a Wikibase instance, installed on Wikimedia Commons to capture linked open data for the over 60 million media files in the media repository. This work is strategically important to the Wikimedia community as it provides the visual source material that drives content creation on all of Wikipedia's sister projects.

When browsing Commons, the main new feature to look for is an added tab where linked open data lives alongside existing free text descriptions (wikitext markup). This inconspicuous upgrade "makes the files on Commons much easier to view, search, edit, curate or organize, use and reuse, in many languages."

Structured Data on Commons presents BHL with a new opportunity to collate its image collection into a central location and normalize over a decade of metadata accretion generated by users and machines.



File:Suíno alta.jpg

From Wikimedia Commons, the free media repository

Size of this preview: 800 × 529 pixels. Other resolutions: 320 × 211 pixels | 640 × 423 pixels | 1,024 ×

Language select: Show all

Original file (3,923 × 2,592 pixels, file size: 2.2 MB, MIME type: image/jpeg); ZoomViewer

Structured data labels link to their respective Wikidata pages and will open in a new window.

Items portrayed in this file

depicts

Figure: The structured data panel for an image in Wikimedia Commons; Swine skeleton, after the technique of bone maceration, on display at the University of São Paulo Museum of Veterinary Anatomy. (Image: Wagner Souza e Silva)

SDC benefits for BHL users are manifold:

- **Greater searchability:** SPARQL-driven queries for BHL images
- **Exposure to Wikimedia ecosystem:** BHL images as source material for all sister projects including Wikipedia’s 300+ language editions.
- **Connected media:** BHL’s images interconnected with other media databases on the web

- **User-generated metadata for BHL images:** CC0, harvestable metadata to be reused by BHL Technical Team and other app developers
- **5-star linked open data**: conversion, normalization, and crowdsourcing of existing image metadata
- **Multi-lingual and accessible data:** BHL images reaching new and under-served audiences through multilingual and accessibility features offered by Wikibase

From the above benefits, perhaps the main advantage for BHL is that the metadata generated by volunteers could be harvested back into the BHL data ecosystem and used to drive native image searching — a long-standing request from BHL's users.

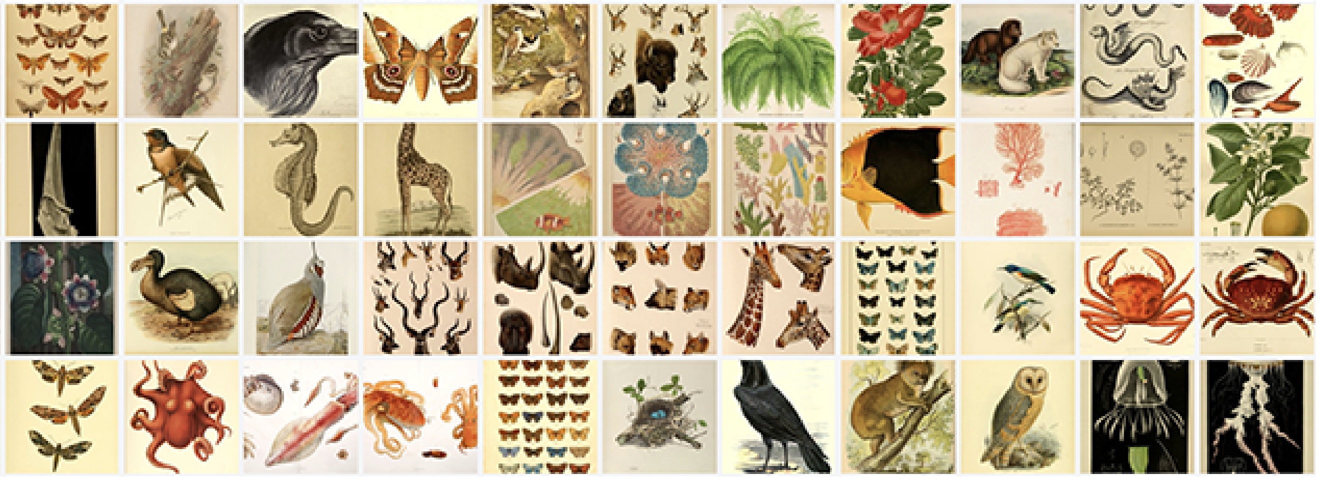


Figure: BHL's contributions to the open access image space have left global audiences awestruck. Over 300,000 scientific images are in the public domain due to the hard work of BHL Staff. (Image: Costantino, 2021)

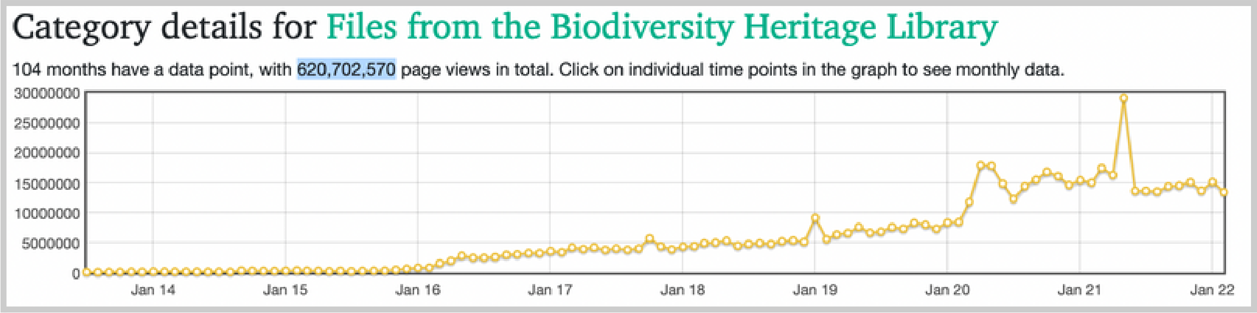
Continuing BHL's image efforts on Wikimedia Commons (SDC) seems to be the logical next step forward for BHL image search – enhanced with 5-star linked open data.

4.2.2 Image Use Metrics and Analytics

Tracking image views and reuse in the Wikimedia ecosystems can give BHL staff insight into the impact and reach of specific campaigns, edit-a-thons, or targeted telework activities. Wikimedia has a suite of community tools that allow BHL to track metrics at the collection, page, and image levels.

BHL Image Use Statistics

BaGLAMa 2 is part of a suite of GLAM metrics tools that allows users to track image statistics by Wikimedia Commons Categories. To date, the 303,302 image files from BHL in Wikimedia Commons have been **viewed over 620 million times!**



Category details for Files from the Biodiversity Heritage Library

104 months have a data point, with 620,702,570 page views in total. Click on individual time points in the graph to see monthly data.

Figure: Growing access – BHL Images have 620,702,570 page views(Image: Dearborn, 2023)

Page-level

BaGLAMa 2 not only tracks collection-level statistics but also can drill down into specific pages where BHL images appear, giving the total view count of the image on a particular page. The tool shows how BHL images strategically placed in a single Wikipedia article can result in thousands of views.

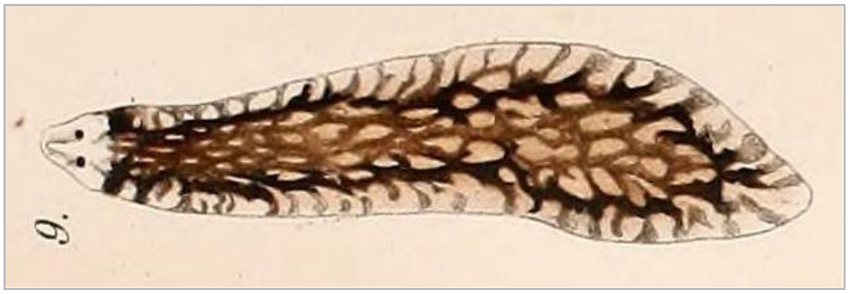


*Figure: Because this image of Sorocelis reticulosa is embedded in the English Wikipedia article for "Animal", it has received **295,023 views** from just that one article alone. (Image: Dearborn, 2023)*

Image-level

In addition to seeing wiki page-level views for an image, one can also see total views on Wikimedia projects for a single image with the **Media Views Analysis** tool. The metrics from this tool allow us to drill down into the source of referral traffic requesting a media file, giving insight into any image's overall proliferation on Wikipedia and other Wikimedia projects.

Figure: This BHL image has been viewed on Wikimedia Projects over 2 million times. (Image: Dearborn, 2023)

While statistics like these are needed to measure the impact of BHL's outreach activities, they aren't everything. Former BHL Communications Manager, Grace Constantino writes that real impact is sometimes not just a numbers game. In March 2019, the #HerNaturalHistory campaign showed that the value of these events is immeasurable:

> *"Although metrics are important, and offer a glimpse into one aspect of these events, the more impactful takeaways from these collaborations were the discussions that attendees had on topics ranging from open content, to information architecture, data storage, content curation, and more. In a sense, these crowd-sourcing edit-a-thons were a salon of sorts, where curious people who are passionate about open information were able to discuss strategies for how to make data more accessible and useful to the public, and work together to find and implement these solutions." (Jackson, E., & Costantino, G., 2019)*

What is most inspiring about Wikimedia events and activities is that they foster both digital and human connections. Metrics can speak to the digital impact BHL media has globally, but the conversations and the spark of curiosity that fosters life-long passions underscores what this knowledge work is all about.

4.3 Future Outlook

Wikimedia Commons is not going anywhere. It has a dedicated community and acts as the image repository for all Wikimedia projects. The community is actively working on the backlog of feature requests. The deployment of the Wikibase extension to store image metadata in Wikimedia Commons was a major milestone for the community, funded by a $3,015,000 grant from the Sloan Foundation.

Like many of the Wikimedia Projects, novice users have noted that the secret rules, hacks, and workarounds coupled with a lack of structure can be overwhelming and intimidating, making navigation and search difficult. (Vershbow, 2018) These barriers to participation have resulted in a relatively small but expert active user community for the project. With outreach efforts around SDC and a new community of SDC data evangelists such as Sandra Fauconnier, the hope is that these former barriers evaporate.

To learn more about Wikimedia Commons, you can refer to this helpful guide.

5.0 Wikisource

5.1 About

Wikisource, is a digital library of freely licensed source texts and historical documents, including poetry, government documents, constitutions of many countries, and general literature. But Wikisource is not just a digital library of open-source texts – it is a free, powerful transcription platform. Originally called "Project Sourceberg," Wikisource was deployed in 2003 to host a small collection of historically significant source texts that underpin article creation in Wikipedia. Today, there are 74 language editions of Wikisource; in the aggregate, there are 988,369 items and 381 active users. (Appendix 3: Statistics)

Wikisource volunteers, sometimes called “Wikisourcers,” are involved in complex librarian-like tasks like cataloging, uploading, proofreading (transcribing), validating, and categorizing historical texts from the public domain.

The ability to proofread, transcribe, and validate texts is driven by the Proofread Page MediaWiki extension.

Wikisource Feature Set:

- **Documentation:** a robust user documentation center
- **Multiple export formats:** EPUB 3, EPUB 2, HTML, MOBI, PDF, RTF, Plain text
- **Statistics:** granular metrics at the project, page, and user-level
- **Version control:** all edits stored as a version with diff capabilities, rollback, and review as needed
- **Language editions:** 74 communities to tap into
- **Interwiki links**: the ability to link titles, authors, pages, words in a text to corresponding items from Wikidata, Wikimedia Commons, and Wikipedia
- **Optical character recognition (OCR) engines:** Tesseract and Google Cloud Vision

Recent BHL OCR Developments

In 2023, BHL Technical Team prioritized its OCR text files in aggregate and placed considerable efforts towards improving the data quality of BHL’s 60-million-page textual corpus. BHL’s Technical Coordinator, Joel Richard, has been running an OCR reprocessing project to improve legacy OCR text using the same open-source OCR engine that Wikisource uses, Tesseract. (Richard, 2022)

For the BHL Technical Team, the OCR reprocessing work is motivated by two main goals:

1. **Surfacing and interlinking scientific species names found in BHL text**

Perhaps BHL’s biggest selling point over other platforms in the biodiversity literature space, is the strength of its collaboration with the Global Names project. Dmitry Mozzherin, Scientific Informatics Leader at Marine Biological Laboratory, and Dr. Geoff Ower, Database Manager at the Catalogue of Life (CoL), are the genius duo behind a suite of web services that:

- find and index biological scientific names found in texts and,
- interconnect found names to over 200 other taxonomic and biodiversity knowledge bases on the web. (Global Names Resolver: Data Sources, n.d.)

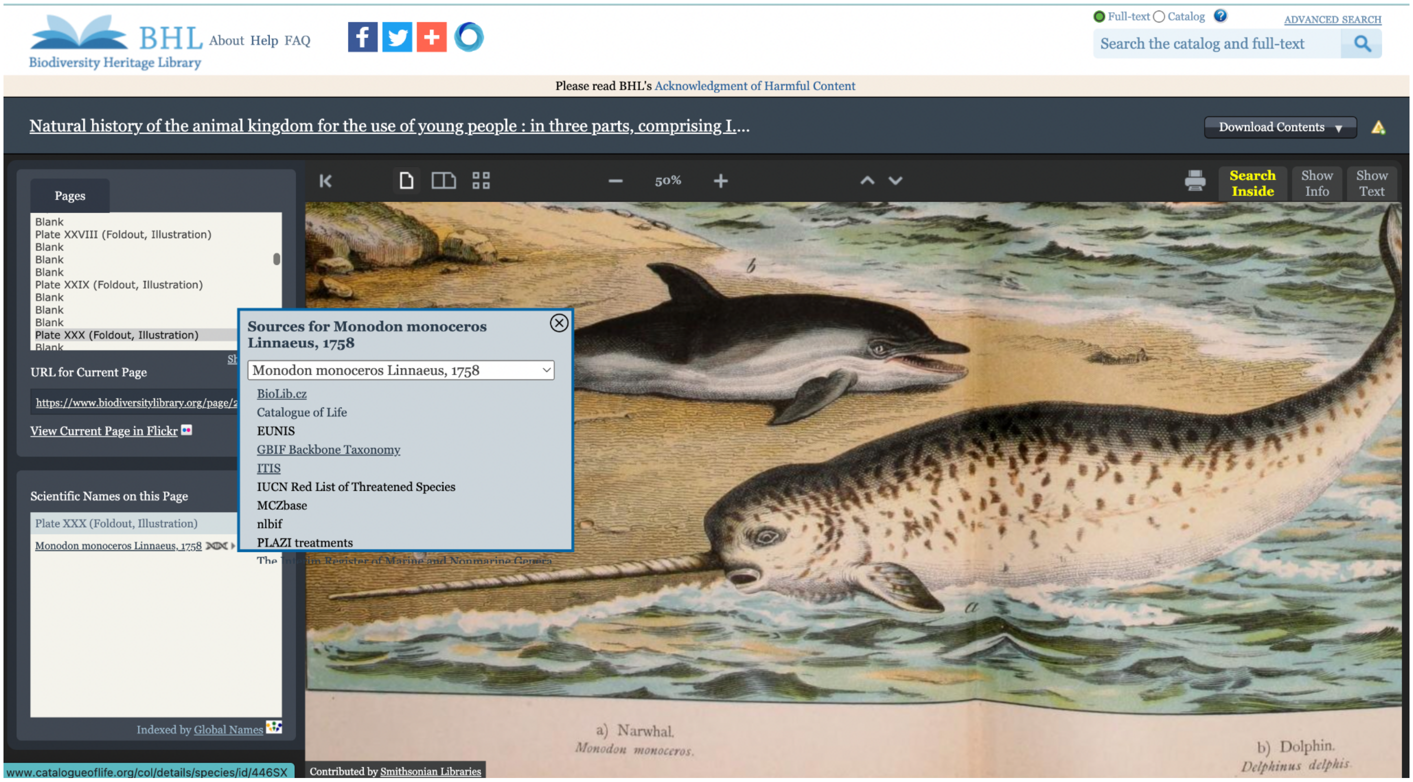
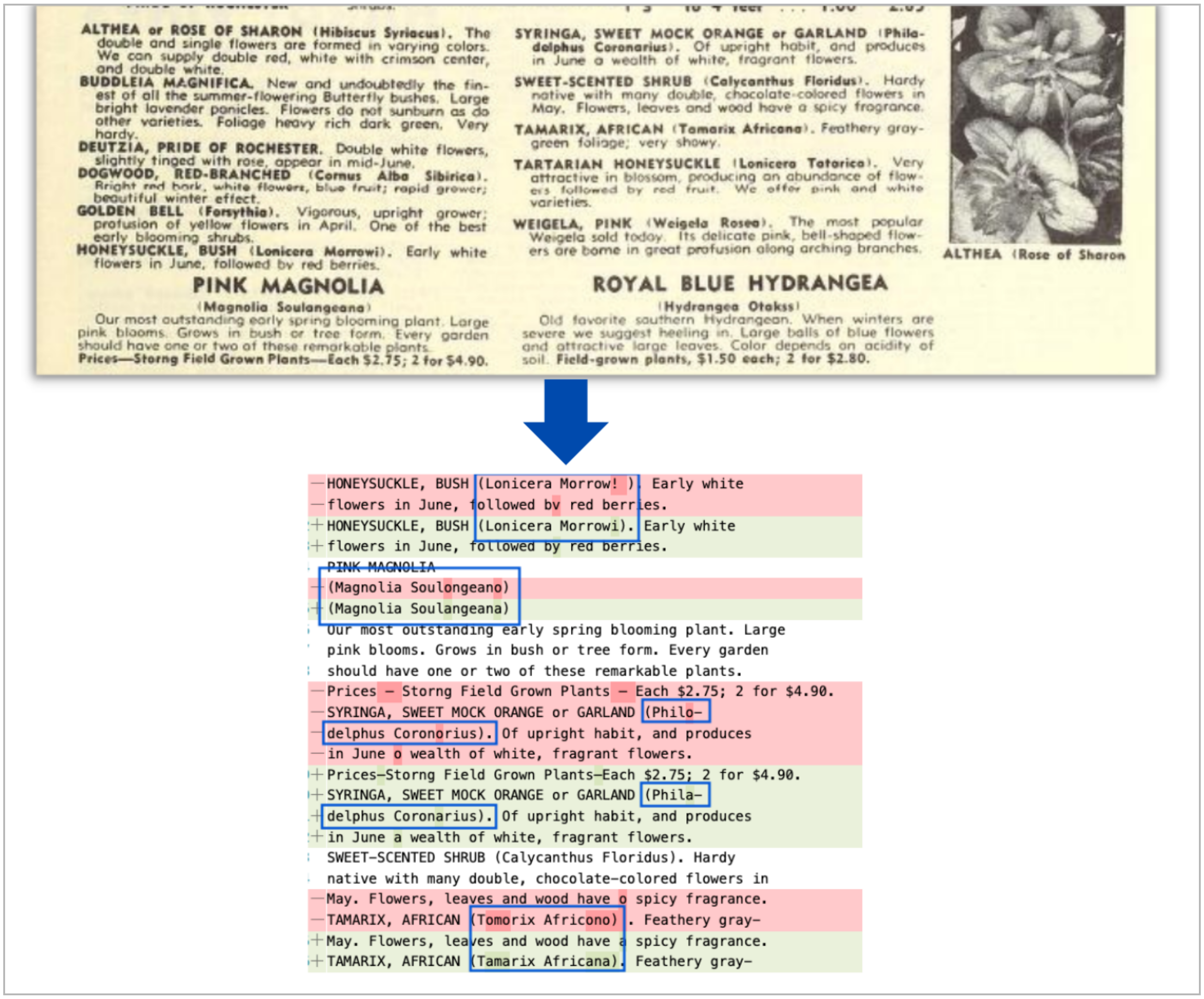

BHL
About Help FAQ
Biodiversity Heritage Library
Full-text Catalog
ADVANCED SEARCH
Search the catalog and full-text
Please read BHL's Acknowledgment of Harmful Content
Natural history of the animal kingdom for the use of young people : in three parts, comprising I....
Download Contents
50%
Search Inside
Show Info
Show Text
Pages
Blank
Plate XXVIII (Foldout, Illustration)
Blank
Blank
Blank
Plate XXIX (Foldout, Illustration)
Blank
Blank
Blank
Plate XXX (Foldout, Illustration)
Blank
URL for Current Page
https://www.biodiversitylibrary.org/page/
View Current Page in Flickr
Scientific Names on this Page
Plate XXX (Foldout, Illustration)
Monodon monoceros Linnaeus, 1758
Indexed by Global Names
Sources for Monodon monoceros Linnaeus, 1758
Monodon monoceros Linnaeus, 1758
BioLib.cz
Catalogue of Life
EUNIS
GBIF Backbone Taxonomy
ITIS
IUCN Red List of Threatened Species
MCZbase
nlbif
PLAZI treatments
The Interim Register of Marine and Nonmarine Genera
b
a
a) Narwhal.
Monodon monoceros.
b) Dolphin.
Delphinus delphis.
www.catalogueoflife.org/col/details/species/id/446SX
Contributed by Smithsonian Libraries

Figure: A search for Monodon monoceros can yield some impressive results in BHL which include media and interlinkages beyond the BHL website to databases like the Catalogue of Life, GBIF Backbone Taxonomy, Integrated Taxonomic Information System (ITIS), and more. Wouldn't it be fantastic to see Wikidata in the list of interconnected databases? (Image: Dearborn, 2023)

In-text species name recognition powered by Global Names' algorithms and APIs has given the BHL Technical Team a major reason to prioritize the quality of BHL's OCR text so that BHL can surface and interlink even more previously hidden scientific names found in the BHL corpus to all the world.



ALTHEA or ROSE OF SHARON (Hibiscus Syriacus). The double and single flowers are formed in varying colors. We can supply double red, white with crimson center, and double white.

BUDDLEIA MAGNIFICA. New and undoubtedly the finest of all the summer-flowering Butterfly bushes. Large bright lavender panicles. Flowers do not sunburn as do other varieties. Foliage heavy rich dark green. Very hardy.

DEUTZIA, PRIDE OF ROCHESTER. Double white flowers, slightly tinged with rose, appear in mid-June.

DOGWOOD, RED-BRANCHED (Cornus Alba Sibirica). Bright red bark, white flowers, blue fruit; rapid grower; beautiful winter effect.

GOLDEN BELL (Forsythia). Vigorous, upright grower; profusion of yellow flowers in April. One of the best early blooming shrubs.

HONEYSUCKLE, BUSH (Lonicera Morrowi). Early white flowers in June, followed by red berries.

PINK MAGNOLIA

(Magnolia Soulangeana)

Our most outstanding early spring blooming plant. Large pink blooms. Grows in bush or tree form. Every garden should have one or two of these remarkable plants.

Prices—Storng Field Grown Plants—Each $2.75; 2 for $4.90.

SYRINGA, SWEET MOCK ORANGE or GARLAND (Philadelphus Coronarius). Of upright habit, and produces in June a wealth of white, fragrant flowers.

SWEET-SCENTED SHRUB (Calycanthus Floridus). Hardy native with many double, chocolate-colored flowers in May. Flowers, leaves and wood have a spicy fragrance.

TAMARIX, AFRICAN (Tamarix Africana). Feathery gray-green foliage; very showy.

TARTARIAN HONEYSUCKLE (Lonicera Tatarica). Very attractive in blossom, producing an abundance of flowers followed by red fruit. We offer pink and white varieties.

WEIGELA, PINK (Weigela Rosea). The most popular Weigela sold today. Its delicate pink, bell-shaped flowers are borne in great profusion along arching branches.

ALTHEA (Rose of Sharon

ROYAL BLUE HYDRANGEA

(Hydrangea Otakss)

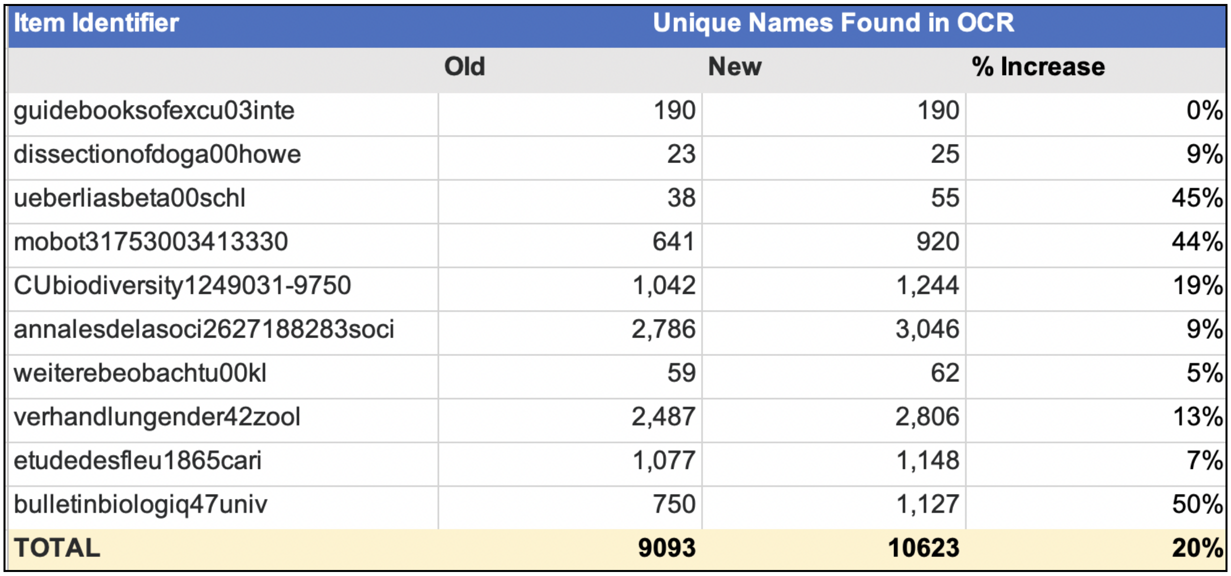
Old favorite southern Hydrangean. When winters are severe we suggest heeling in. Large balls of blue flowers and attractive large leaves. Color depends on acidity of soil. **Field-grown plants, $1.50 each; 2 for $2.80.**

```
—HONEYSUCKLE, BUSH (Lonicera Morrow! ). Early white
—flowers in June, followed bv red berries.
+HONEYSUCKLE, BUSH (Lonicera Morrowi). Early white
+flowers in June, followed by red berries.
 PINK MAGNOLIA
—(Magnolia Soulongeano)
+(Magnolia Soulangeana)
 Our most outstanding early spring blooming plant. Large
 pink blooms. Grows in bush or tree form. Every garden
 should have one or two of these remarkable plants.
—Prices — Storng Field Grown Plants — Each $2.75; 2 for $4.90.
—SYRINGA, SWEET MOCK ORANGE or GARLAND (Philo-
—delphus Coronorius). Of upright habit, and produces
—in June o wealth of white, fragrant flowers.
+Prices—Storng Field Grown Plants—Each $2.75; 2 for $4.90.
+SYRINGA, SWEET MOCK ORANGE or GARLAND (Phila-
+delphus Coronarius). Of upright habit, and produces
+in June a wealth of white, fragrant flowers.
 SWEET-SCENTED SHRUB (Calycanthus Floridus). Hardy
 native with many double, chocolate-colored flowers in
—May. Flowers, leaves and wood have o spicy fragrance.
—TAMARIX, AFRICAN (Tomorix Africono) . Feathery gray-
+May. Flowers, leaves and wood have a spicy fragrance.
+TAMARIX, AFRICAN (Tamarix Africana). Feathery gray-
```

Figure: A snippet of a single page in BHL and its OCR outputs. The side-by-side comparison shows scientific names are more accurately recognized in Tesseract OCR engine outputs. This impacts the quality of BHL's species bibliographies and its ability to interlink with biological databases in the BHL scientific name search feature. (Richard, 2022)

2. Improving text output accuracy with rapidly improving technology in the OCR engine space

BHL's OCR reprocessing project is working through 120,000 legacy OCR files to improve its OCR text outputs which thereby will surface more scientific names found in the text while also improving BHL's full-text search functionality for other named entity searches.



Item Identifier	Unique Names Found in OCR		
	Old	New	% Increase
guidebooksofexcu03inte	190	190	0%
dissectionofdoga00howe	23	25	9%
ueberliasbeta00schl	38	55	45%
mobot31753003413330	641	920	44%
CUbiodiversity1249031-9750	1,042	1,244	19%
annalesdelasoci2627188283soci	2,786	3,046	9%
weiterebeobachtu00kl	59	62	5%
verhandlungender42zool	2,487	2,806	13%
etudedesfleu1865cari	1,077	1,148	7%
bulletinbiologiq47univ	750	1,127	50%
TOTAL	**9093**	**10623**	**20%**

Figure: Preliminary results of a small sample set show that by reprocessing BHL's legacy OCR, the BHL Tech Team will likely improve Global Names name-finding algorithm matches by an average of 20%. (Richard, 2022)

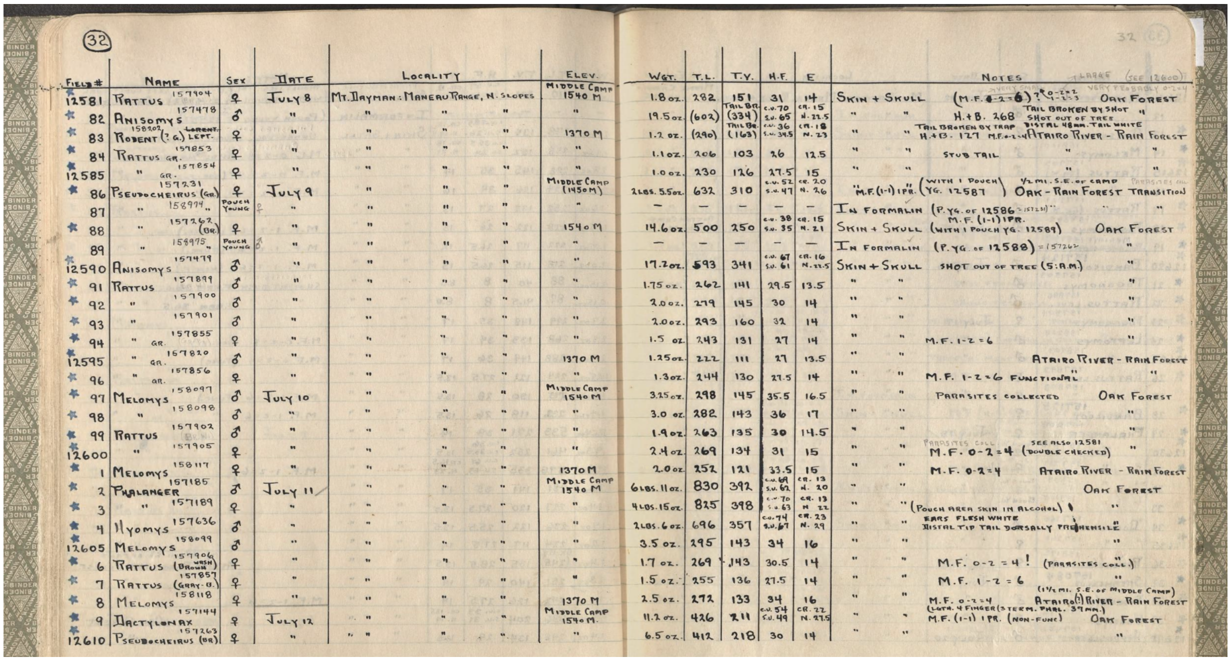
Despite these exciting improvements that improve the accuracy of BHL's OCR text files, the BHL Technical Team estimates the project will take nearly two years to complete due to the processing power and computational resources required. Additionally, BHL's Transcription Upload Tool Working Group (BHL-TUTWG) reports that the sub-corpus of handwritten content in BHL is roughly one million pages. This may seem like a small percentage of the corpus. However, it includes BHL's collection of scientific field notes, which are rich sources of taxonomic names, species occurrence records, and ecological, geolocation, climate, and weather data.

Unfortunately, archival, handwritten, and incunabula content will likely benefit much less from the BHL OCR reprocessing work. Instead, more technical resources to surface species names, climate, and biological data in these materials are required. A few examples of data that could benefit from being processed by a Handwriting Text Recognition (HTR) engine found in BHL's Scientific Field Notes Collection:

- Hydrographic data
- Weather data
- Plant lists
- Species distribution maps
- Sampling event data

- Checklist data
- Specimen photos

Further extraction of this data could be used to directly support conservation and climate change research. Katie Mika, Data Services Librarian at Harvard University and former BHL National Digital Stewardship Resident, has been collaborating with the BHL Technical Team to pilot a data extraction pipeline aimed at freeing this valuable data from BHL's field notes. (Dearborn & Mika, 2022)(Appendix 2: Interviews)



32

Field #	Name	Sex	Date	Locality	Elev.	Wgt.	T.L.	T.V.	H.F.	E	Notes
12581	157904 Rattus	♀	July 8	Mt. Dayman: Maneau Range, N. slopes	Middle Camp 1540 M	1.8 oz.	282	151	31	14	Skin + Skull (M.F. ~~0-2=6~~)? very small 0-2=2 1-2=3 very probably 0-2=4 large (see 12600) Oak Forest
82	157478 Anisomys	♂	"	" " " "	"	19.5 oz.	(602)	Tail Br. (334)	c.v. 70 s.v. 65	cr. 15 N. 22.5	" " H.+B. 268 Tail broken by shot Shot out of tree "
83	158202 ~~Lorent.~~ Robent (? G.) Lept.	♀	"	" " " "	1370 M	1.2 oz.	(290)	Tail Br. (163)	c.v. 36 s.v. 34.5	cr. 18 N. 23	" " Tail broken by trap H.+B. 127 M.F. 0-2=4 Distal 48mm. tail white Atairo River - Rain Forest
84	157853 Rattus gr.	♀	"	" " " "	"	1.1 oz.	206	103	26	12.5	" " Stub tail " "
12585	157854 " gr.	♀	"	" " " "	"	1.0 oz.	230	126	27.5	15	" " " "
86	157231 Pseudocheirus (gr.)	♀	July 9	" " " "	Middle Camp (1450 M)	2 lbs. 5.5 oz.	632	310	c.v. 52 s.v. 47	cr. 20 N. 26	" M.F. (1-1) 1 pr. " (With 1 pouch yg. 12587) ½ mi. S.E. of camp Oak - Rain Forest Transition Parasites coll.
87	158974 " "	Pouch young ♀	"	" " " "	"	—	—	—	—	—	In formalin (P. yg. of 12586 = 157231) " "
88	157262 " (Br.)	♀	"	" " " "	1540 M	14.6 oz.	500	250	c.v. 38 s.v. 35	cr. 15 N. 21	Skin + Skull M.F. (1-1) 1 pr. (with 1 pouch yg. 12589) Oak Forest
89	158975 " "	Pouch young ♂	"	" " " "	"	—	—	—	—	—	In formalin (P. yg. of 12588) = 157262 "
12590	157479 Anisomys	♂	"	" " " "	"	17.2 oz.	593	341	c.v. 67 s.v. 61	cr. 16 N. 22.5	Skin + Skull shot out of tree (5:A.M.) "
91	157899 Rattus	♂	"	" " " "	"	1.75 oz.	262	141	29.5	13.5	" " "
92	157900 "	♂	"	" " " "	"	2.0 oz.	279	145	30	14	" " "
93	157901 "	♂	"	" " " "	"	2.0 oz.	293	160	32	14	" " "
94	157855 " gr.	♀	"	" " " "	"	1.5 oz.	243	131	27	14	" " M.F. 1-2=6 "
12595	157820 " gr.	♂	"	" " " "	1370 M	1.25 oz.	222	111	27	13.5	" " Atairo River - Rain Forest
96	157856 " gr.	♀	"	" " " "	"	1.3 oz.	244	130	27.5	14	" " M.F. 1-2=6 functional " "
97	158097 Melomys	♂	July 10	" " " "	Middle Camp 1540 M	3.25 oz.	298	145	35.5	16.5	" " Parasites collected Oak Forest
98	158098 "	♂	"	" " " "	"	3.0 oz.	282	143	36	17	" " "
99	157902 Rattus	♂	"	" " " "	"	1.9 oz.	263	135	30	14.5	" " "
12600	157905 "	♀	"	" " " "	"	2.4 oz.	269	134	31	15	" " Parasites coll. M.F. 0-2=4 (double checked) See also 12581 "
1	158117 Melomys	♀	"	" " " "	1370 M	2.0 oz.	252	121	33.5	15	" " M.F. 0-2=4 Atairo River - Rain Forest
2	157185 Phalanger	♂	July 11	" " " "	Middle Camp 1540 M	6 lbs. 11 oz.	830	392	c.v. 69 s.v. 62	cr. 13 N. 20	" " Oak Forest
3	157189 "	♀	"	" " " "	"	4 lbs. 15 oz.	825	398	c.v. 70 s.v. 63	cr. 13 N 22	" " (Pouch area skin in alcohol) "
4	157636 Hyomys	♂	"	" " " "	"	2 lbs. 6 oz.	696	357	c.v. 74 s.v. 67	cr. 23 N. 29	" " Ears flesh white Distal tip tail dorsally prehensile "
12605	158099 Melomys	♂	"	" " " "	"	3.5 oz.	295	143	34	16	" " "
6	157906 Rattus (Brown wash)	♀	"	" " " "	"	1.7 oz.	269	143	30.5	14	" " M.F. 0-2=4! (Parasites coll.)
7	157857 Rattus (Gray-B.)	♀	"	" " " "	"	1.5 oz.	255	136	27.5	14	" " M.F. 1-2=6 "
8	158118 Melomys	♀	"	" " " "	1370 M	2.5 oz.	272	133	34	16	" " M.F. 0-2=4 (1½ mi. S.E. of Middle Camp) Atairo(?) River - Rain Forest
9	157144 Dactylonax	♀	July 12	" " " "	Middle Camp 1540 M.	11.2 oz.	426	211	c.v. 54 s.v. 49	cr. 22 N. 27.5	" " (Lgth. 4 finger (3 term. phal. 39 mm.) M.F. (1-1) 1 pr. (Non-func) Oak Forest
12610	157263 Pseudocheirus (Br)	♀	"	" " " "	"	6.5 oz.	412	218	30	14	" " "

Figure: Data in field books include high-value biodiversity data that will require HTR engine processing and text extraction techniques to make the data usable. In addition to the catalog entries, recapitulated lists of data sorted by animal type, date of collection, and type of specimen, and notes on languages and localities. (Image: American Museum of Natural History, BHL item 229794)

5.2 Identified Use Cases

5.2.2 Handwritten Text Transcription

To get at this impactful data, Lead Developer Mike Lichtenberg and BHL's Transcription Upload Tool Working Group have been testing the accuracy of three cutting-edge HTR engines that can handle a multitude of pre-modern type scripts and human handwriting styles:

- Microsoft Azure,
- Amazon Textract,
- Google Cloud Vision.

Text engines have come a long way just in the last few years. HTR Engines are deploying machine learning techniques that are finally allowing BHL to extract OCR from handwritten texts, incunabula, and early manuscript black letter typescripts like Fraktur.

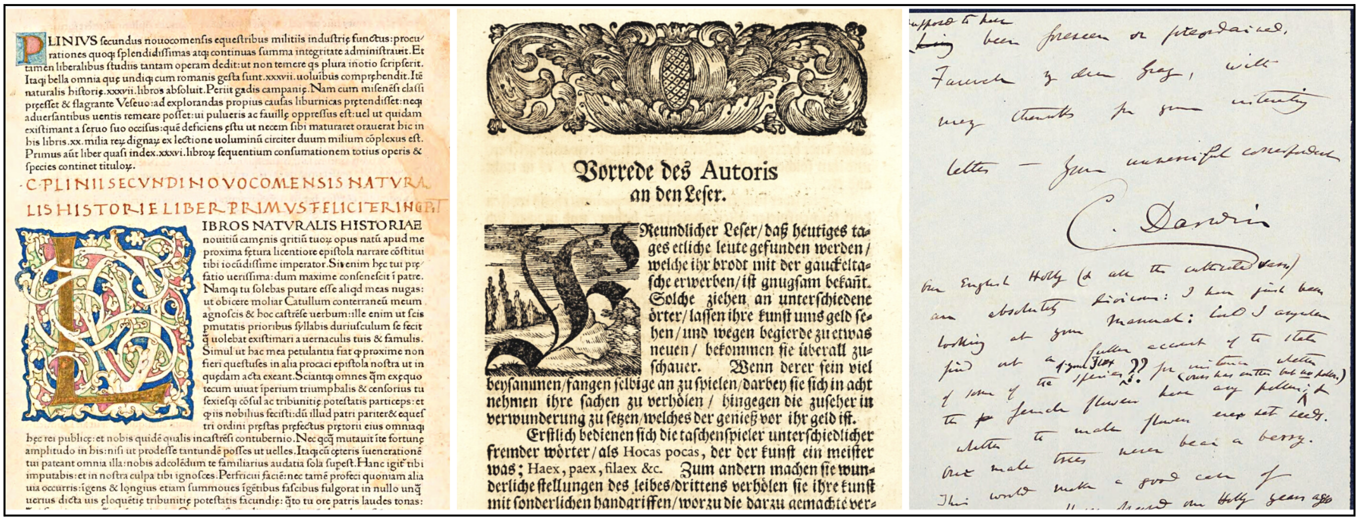
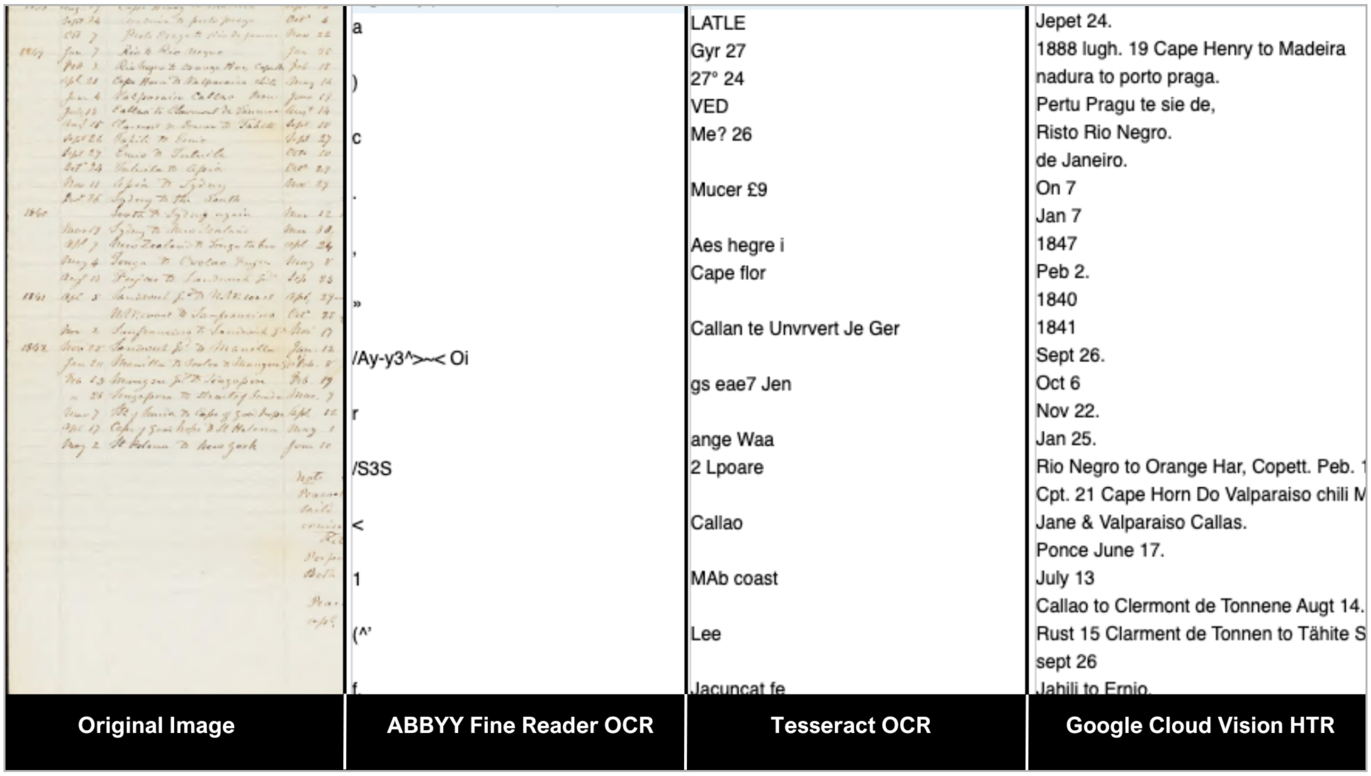


Figure: Incunabula, Fraktur, and handwriting present additional challenges that will need to be overcome if the desire is to have a full-text search for all of the BHL corpus. (Image: Dearborn, 2023)

A sticking point with these cutting-edge services for BHL is that they have a cost attached and it needs to be determined whether the outputs are “good enough” to justify the cost for the Consortium.

While these experiments are in progress, BHL partner institutions can use, free-of-charge, Google Cloud Vision by way of Wikisource. Wikisource, which has deployed the Wikimedia OCR extension, actually offers GLAM communities *both* Tesseract OCR and Google Cloud Vision HTR. For institutions that would like to hasten the transcription work of digitized archival materials in their collections, or simply improve the OCR for published works using older typescripts, Wikisource indeed provides.



ABBYY Fine Reader OCR

```
a
)
c
.
,
»
/Ay-y3^>~< Oi
r
/S3S
<
1
(^'
f
```

Tesseract OCR

```
LATLE
Gyr 27
27° 24
VED
Me? 26

Mucer £9

Aes hegre i
Cape flor

Callan te Unvrvert Je Ger

gs eae7 Jen

ange Waa
2 Lpoare

Callao

MAb coast

Lee

Jacuncat fe
```

Google Cloud Vision HTR

```
Jepet 24.
1888 lugh. 19 Cape Henry to Madeira
nadura to porto praga.
Pertu Pragu te sie de,
Risto Rio Negro.
de Janeiro.
On 7
Jan 7
1847
Peb 2.
1840
1841
Sept 26.
Oct 6
Nov 22.
Jan 25.
Rio Negro to Orange Har, Copett. Peb. 1
Cpt. 21 Cape Horn Do Valparaiso chili M
Jane & Valparaiso Callas.
Ponce June 17.
July 13
Callao to Clermont de Tonnene Augt 14.
Rust 15 Clarment de Tonnen to Tähite S
sept 26
Jahili to Ernio.
```

Figure: A side-by-side comparison of outputs via various text engines in Wikisource. HTR isn't perfect by any means but certainly comes a long way in delivering better data. The results will only improve as HTR engines advance. (Image: Dearborn, 2023)

5.3 Future Outlook

BHL Staff have been watching Wikisource for a long time after having met Andrea Zanni, former Wikisource sysop and President of Wikimedia Italia, at Wikimania 2012. He envisioned a library of the future which he called the "hyper-library" – where texts interlink with other texts, other Wikimedia sister projects, and the broader Web. This interlinking allows us to collectively create a virtual library space where the serendipitous discovery of new knowledge happens.

To make Andrea's vision come to life and make Wikisource the go-to GLAM transcription platform, Wikimedia needs to make further investments in Wikisource.

It is important to note that Wikisource faces steep competition. There are already many platforms for text transcription and choosing the right one is already a daunting, often paralyzing decision for capacity-constrained GLAMs. Compared to other transcription software available on the market, Wikisource's user interface feels clunky and crowded. However, for what it lacks in the aesthetics department, it does make up for it with a powerful feature set.

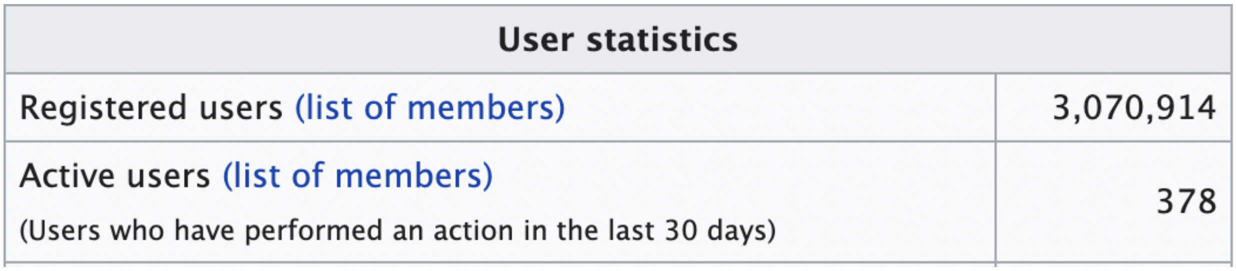
Wikisource's strategic investment should focus on five key areas:

1. Overhauling its user-Interface,
2. Creating a GLAM project-based tool kit,
3. Building a structured data extraction extension,
4. Improving the Wikisource API and/or OAI-PMH feed to facilitate text ingestion between databases, and
5. Improving integration with Wikidata.

Underpinning these key areas is the Wikisource community. The Wikimedia Foundation will need to do work to galvanize, retain, and build the numbers and strength of the Wikisource community which are incredibly sophisticated and underutilized. Without needing to be semantic web experts, Wikisourcers already intuitively understand interlinking's impact on knowledge discovery:

> *"the power and potential of [Wikisource] is mind-blowing": where one work refers to another (via wikilinking), thereby contributing to the "great conversation." One author, one book leads to another author, another book—another idea." (Bohm, 2016)*

Compared to other sister projects, Wikisource exhibits a much higher attrition rate.

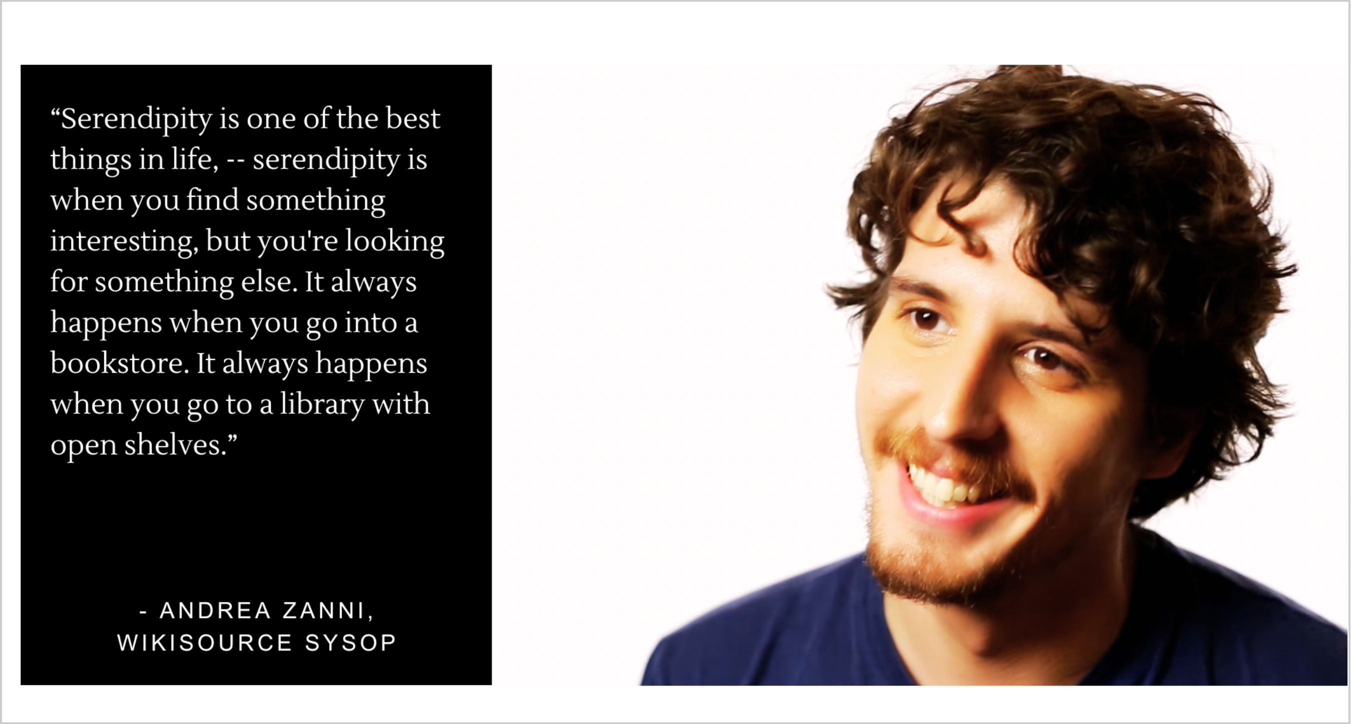


User statistics	
Registered users (list of members)	3,070,914
Active users (list of members) (Users who have performed an action in the last 30 days)	378

Figure: Retaining an active user base should be at the forefront of strategic planning work for Wikisource. (Image: Wikisource Statistics)

BHL has a lot of homework to do as well. BHL must broaden its perspective on OCR text files. This data is not just a driver of full-text search on the BHL platform or something that makes the BHL book viewer more useful. BHL's OCR text in the aggregate is BHL's greatest asset: 500+ years of human scientific knowledge as a 60 million page dataset. The untapped potential of this dataset boggles the mind. To make it useful to computational researchers and proliferate its contents on the web, BHL needs to continue to invest in:

1. Improving overall text recognition accuracy,
2. Hastening the upload of already transcribed materials from BHL partners,
3. Extracting tabular data and depositing it to impactful knowledge bases, and
4. Exploring new features and ways to interlink named entities in the text internally and externally to other related resources.



"Serendipity is one of the best things in life, -- serendipity is when you find something interesting, but you're looking for something else. It always happens when you go into a bookstore. It always happens when you go to a library with open shelves."

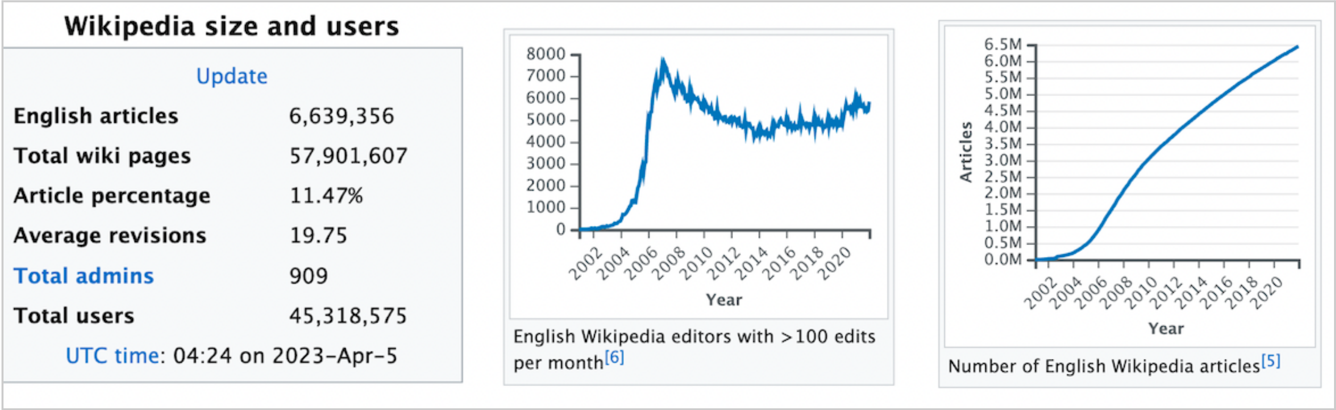
- ANDREA ZANNI,
WIKISOURCE SYSOP

Figure: Andrea Zanni on creating a connected integrated digital library with Wikisource and how wiki interlinking fosters serendipitous knowledge discovery. (Image: Wikimedia Foundation, 2012)

6.0 Wikipedia

6.1 About

Wikipedia is a behemoth, exercising massive authority across the web. The site consistently ranks in the top 10 most visited sites in the world, receiving an average of 4+ billion site visits per month.



Wikipedia size and users

Update

English articles	6,639,356
Total wiki pages	57,901,607
Article percentage	11.47%
Average revisions	19.75
Total admins	909
Total users	45,318,575

UTC time: 04:24 on 2023-Apr-5

English Wikipedia editors with >100 edits per month[6]

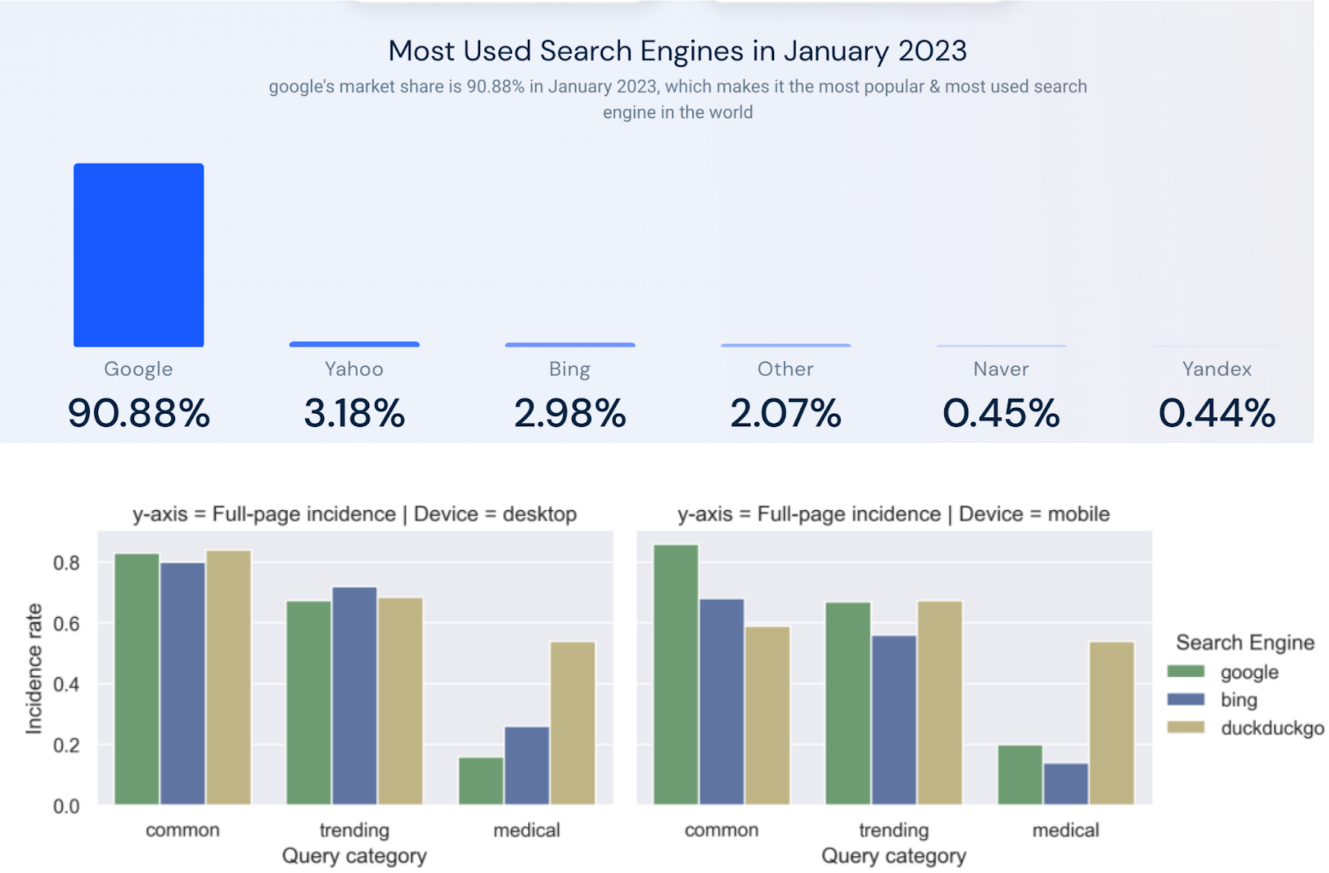
Number of English Wikipedia articles[5]

Figure: Wikipedia is a growing knowledge repository with incredible clout and influence on the web. Additionally, these stats only account for English Wikipedia; there are 332 language editions of Wikipedia and let's not forget the myriad of sister projects. (Image: Wikipedians, 2023)

Wikipedia is now a go-to source for fact-checking (O'Neil, 2021) and many academics argue that Wikipedia is more reliable than a scholarly peer-reviewed article:

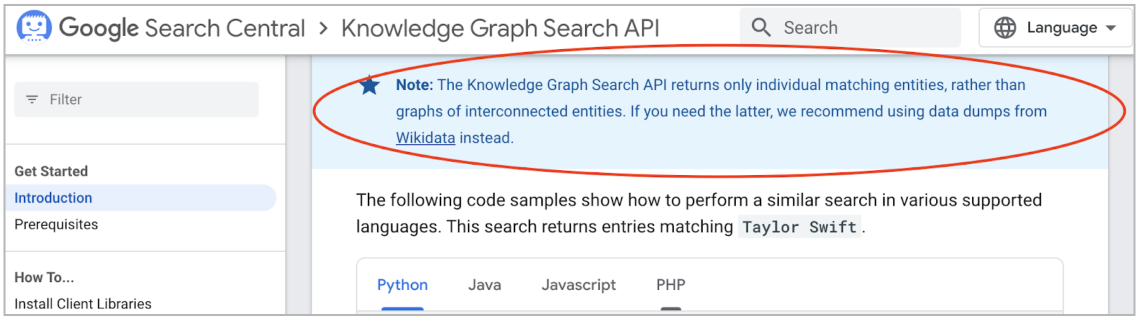
> *"...I argue that the content of a popular Wikipedia page is actually the most reliable form of information ever created. Think about it—a peer-reviewed journal article is reviewed by three experts (who may or may not actually check every detail), and then is set in stone. The contents of a popular Wikipedia page might be reviewed by thousands of people." (Stuart, 2021)*

Perhaps, a lesser-known fact about Wikipedia is that a majority of Wikipedia's impressive user traffic comes from its symbiotic relationship with Google. Wikimedia's Diff Blog purports that "75% of reader sessions (pages viewed by the same user) come from search engines and 90% of these come from a single search engine, Google.” (Johnson et. al., 2021)



Figures: Google’s Search Engine market share accounts for 90% of all searches on the web and Wikipedia appears on page one of common results 80-84% of the time on both desktop and mobile devices. (Images: Top – Similarweb, 2023; Bottom - Vincent & Hecht, 2021)

These impressive stats are due in part to the fact that Google decommissioned its knowledge graph database in 2014 (Pellissier Tanon, et.al., 2016) and now uses Wikipedia and Wikidata as pre-eminent sources of structured data to drive its Knowledge Graph Search API.



Note: The Knowledge Graph Search API returns only individual matching entities, rather than graphs of interconnected entities. If you need the latter, we recommend using data dumps from Wikidata instead.

The following code samples show how to perform a similar search in various supported languages. This search returns entries matching `Taylor Swift`.

Figure: Google Search APIs rely on a myriad of data sources, but a note on the developer page of the Google Knowledge Graph Search API explicitly calls out Wikidata as a primary source of data. (Image: Google, n.d.)

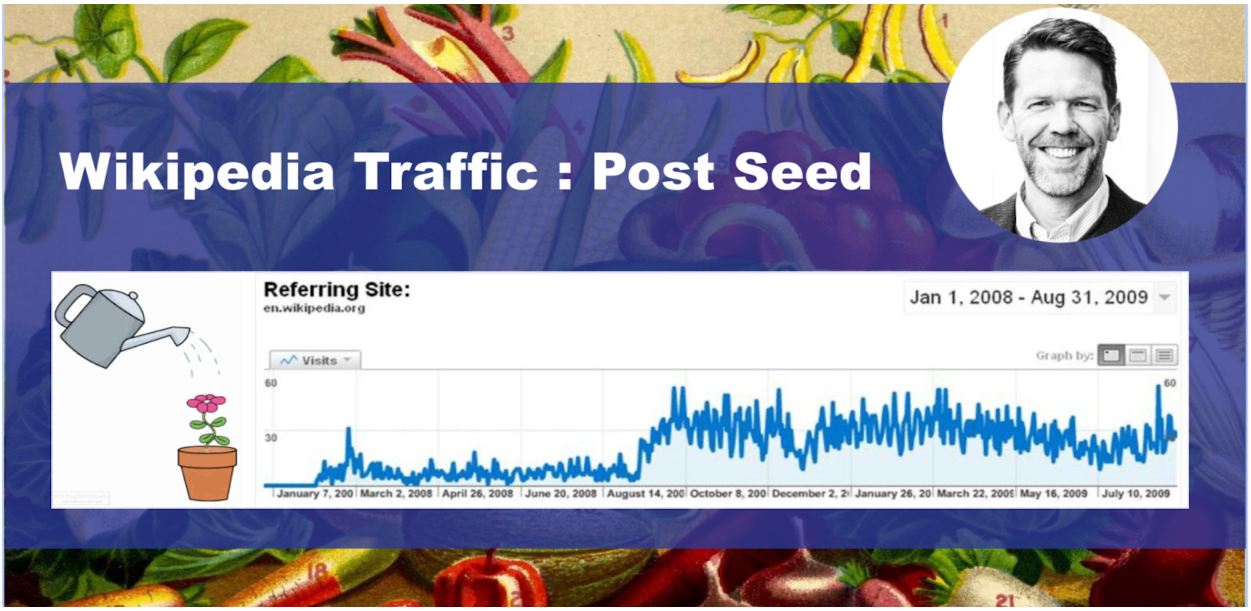
Siobhan Leachman explains how this all comes together in a "virtuous circle of reuse" and it isn't just Google that uses Wikipedia and backbone sister projects to proliferate BHL content throughout the web, it's other biodiversity knowledge bases as well:

> *"But it isn't only Google that ingests Wikipedia, Wikidata and Wikimedia Commons content. Citizen science websites such as iNaturalist also use them. After writing an article in Wikipedia, that article can then be found in iNaturalist. If you go to the species page and look at the "About" tab you will see a copy of the Wikipedia species article. Wikimedia Commons images can also be curated into iNaturalist. [...] It assists citizen scientists in identifying their current biodiversity observations. Once an observation is uploaded and confirmed in iNaturalist, the observation data is in turn ingested into GBIF (Global Biodiversity Information Facility). So BHL plays a part in ensuring the improved accuracy of data generated by iNaturalist and used in GBIF." (Leachman, 2018)*

6.2 Identified Use Cases

6.2.1 Increasing User Traffic to BHL

Intuiting Wikipedia's growing clout on the web, in 2012 Chris Freeland, BHL's former Technical Director decided to run an experiment to seed BHL links in Wikipedia and measure any difference in traffic on those seeded links. His experiment confirmed that seeded links in Wikipedia grow BHL's user traffic. (Freeland, 2012) For BHL staff, this began to illume the idea that the open access journey does not end with digitization but rather it marks the beginning. To truly have an impact on the broader web, content can not stay siloed in BHL's databases.



Wikipedia Traffic : Post Seed

Referring Site:
en.wikipedia.org

Jan 1, 2008 - Aug 31, 2009

Figure: Growing access into reuse; BHL's data journey extends beyond its databases and Wikipedia is a pathway to the broader web. (Image: Dearborn, J., & Kalfatovic, M. 2022)

6.2.2 Bridging Knowledge Gaps with Campaigns

Continuing with Freeland's findings, Grace Costantino, BHL's former Communications Manager, spearheaded content-creation efforts in Wikipedia through multifaceted campaign strategies produced in collaboration with BHL's global partners to further BHL's strategic objectives to fill knowledge gaps. Themes included species conservation, surfacing women in science, and exposing natural history collections as data for global consumption.

To expose women in science, Grace found an enthusiastic Wikimedia ally in Siobhan Leachman, who was one of many who contributed to and helped coordinate the Wikipedia elements of sophisticated campaigns to raise the notability of women in science on the web. These *Her Natural History* campaigns have been incredibly successful, employing such tactics as:

- Persistent identifier generation
- Scholarly blogs by BHL partner staff

- Finding and adding images to Flickr of women in the BHL corpus
- Hosting edit-a-thons
- Campaign promotion through social media channels

Wikipedia Editing Workshop *for* Women in Natural History

When? 13 March 2019. 10am-2pm ET.
Where? Smithsonian Libraries & Remotely.

83
91
BHL
WIKIMEDIA
DISTRICT OF COLUMBIA
Smithsonian
Libraries

The illustrations featured in this graphic were created by women who will be featured in our Wikipedia Editing Workshop. From left to right: Harriet Scott Morgan, *Australian lepidoptera and their transformations*, v. 1 (1864); Roberta McIntosh, *A monograph of the British marine annelids*, v.2, pt.2 (1910); Elizabeth Twining, *Illustrations of the natural orders of plants with groups and descriptions*, v. 2 (1868); Mary Emily Eaton, *Field book of common gilled mushrooms* (1928); Elizabeth Gould, *The Birds of Europe*, v. 1 (1837).

New data points generated by BHL staff for these campaigns in the form of identifiers and blogs, would coalesce into a single Wikidata item for each woman. Once a Wikidata item is established, it provides the requisite notability criteria and source material from which a Wikipedia article can be written (Leachman, 2019) In terms of workflow, Wikidata establishes "facts" (i.e. referenced claims) and Wikipedia delivers them for human consumption. Siobhan's workflows serve as models that could be adopted for other underrepresented groups.

6.3 Future Outlook

Since 2008, BHL has been aware that Wikipedia is a logical pathway towards de-siloing BHL's content. Yet despite all of the beneficial outcomes Wikipedia campaigns bring for GLAMs, let's face it – article creation in Wikipedia is hard. To be a successful Wikipedia editor, one must master a maze of rules which undergird the website, including but not limited to the following policies: Wikipedia's Five Pillars, Wikipedia's Manual of Style, quality of the article.

These policies are there for a good reason as necessary measures in the fight against disinformation campaigns which will only grow exponentially in the coming years with the rise of AI. Wikipedia is no stranger to internet attacks on its content and must remain vigilant in defense of knowledge integrity. (Wikipedia Contributors, 2023) Now more than ever, we need evidence-based discourse.

Nevertheless, there remain knowledge equity problems to solve. The Wikimedia Research Team recognizes that vast and important knowledge domains have been omitted from Wikipedia due to restrictive content policies currently in place:

"While Wikipedia's sourcing policies prevent information lacking reliable secondary sources from being considered, these policies have also prevented vast and important domains of knowledge from entering the project. This is particularly critical for cultures that rely on means of knowledge transmission and representation such as oral sources."

To understand knowledge domain gaps, Alex Stinson, Lead Strategist at the Wikimedia Foundation, classifies them into three measurable areas (each seemingly more difficult to tackle than the last)

1. Coverage - Content created for or about any given knowledge entity
2. Language - Content translated into all the world's languages
3. Geography - Content sourced from and/or about a specific region in the world

(Appendix 2: Interviews)

For a comprehensive look at knowledge gaps, the Wikimedia Research Team has developed a *knowledge gaps taxonomy* to help evaluate and measure inequalities that may exist in terms of knowledge creation and access. The ultimate goal of this work is to create a visualization tool or dashboard that will allow Wikimedians to see where the knowledge gaps exist in real time.

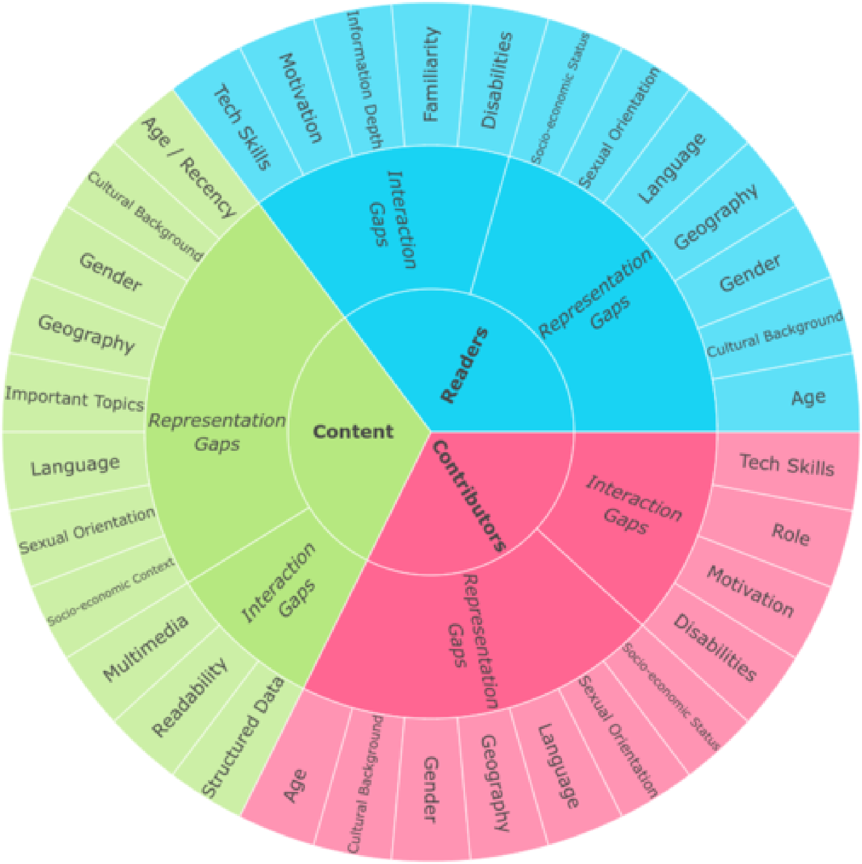


Figure: The Wikimedia Research Team's Knowledge Gap Taxonomy is incredibly helpful to all GLAMs in their prioritization of content and data delivery efforts.

As GLAM professionals, we know there are wide disparities across these areas, and much of our work has been focused on solving this immense problem. But efforts to illuminate all human knowledge will need to scale rapidly to have any real impact on the underserved and underrepresented groups that are now being most threatened by the ravages of climate change. Many of these groups, which possess vast stores of biodiversity knowledge, transmit knowledge through oral traditions, not written ones. Indigenous communities and cultures and their knowledge about our natural world remain largely unaccounted for. Despite positive developments on the horizon for recognizing the unique contributions of Indigenous Knowledge for public policy, especially in regards to our relationship with nature, actionable steps need to be identified and taken (White House, 2022). Will Wikipedia community sysops review their policies to allow other knowledge creators to participate and be represented in the sum of all human knowledge? One hopes.

> *“Because the stories we tell—whether about climate change, United States history, cultural revitalization, or numerous other subjects—matter. Stories shape how we see one another, how we understand our pasts and presents, and how we collectively shape our futures. “ (Keeler, K., 2023)*

7.0 Conclusion

The BHL community has already come together to build the world’s largest open access digital library for biodiversity literature and archives. BHL is a vital knowledge repository containing over 500 years of media, text, and data, about Life on Earth. In the process of liberating content from physical materials, we have learned more about how valuable the data, born in literature, can be to supporting life on a sustainable planet.

BHL’s value should not only be measured by the content it has to offer the world. BHL’s global collective of naturalists, academic scholars, scientific researchers, librarians, information architects, and web developers represent a community that is diverse, deeply knowledgeable, and committed to BHL’s goal to provide “free, worldwide access to knowledge about life on Earth.” (BHL About, n.d.) BHL Staff are *both* data *curators* and *generators*; creating metadata, imaging collections, and sharing untold stories on the web. Together, with Wikimedians, universal bioliteracy isn’t just a dream, it can become our shared reality.

The need for the global biodiversity community and its disparate data silos to unify and build a biodiversity knowledge graph rich in human and machine-curated interlinkages has never been greater. It is the missing “technical infrastructure” sought after by climate policymakers, national governments, and intergovernmental organizations. Converting BHL’s dark data into 5-star linked open data that can be shared widely, improved upon, and made actionable is BHL’s current challenge.

> *“If we want to address the big challenges we face around the future of land use, conservation, climate change, food security, and health, we need efficient ways to bring together all the data capable of helping us understand the changing state of the world and the essential role*

that biodiversity plays at all scales." Donald Hobern, Former Executive Secretary of GBIF. (GBIF, 2018)

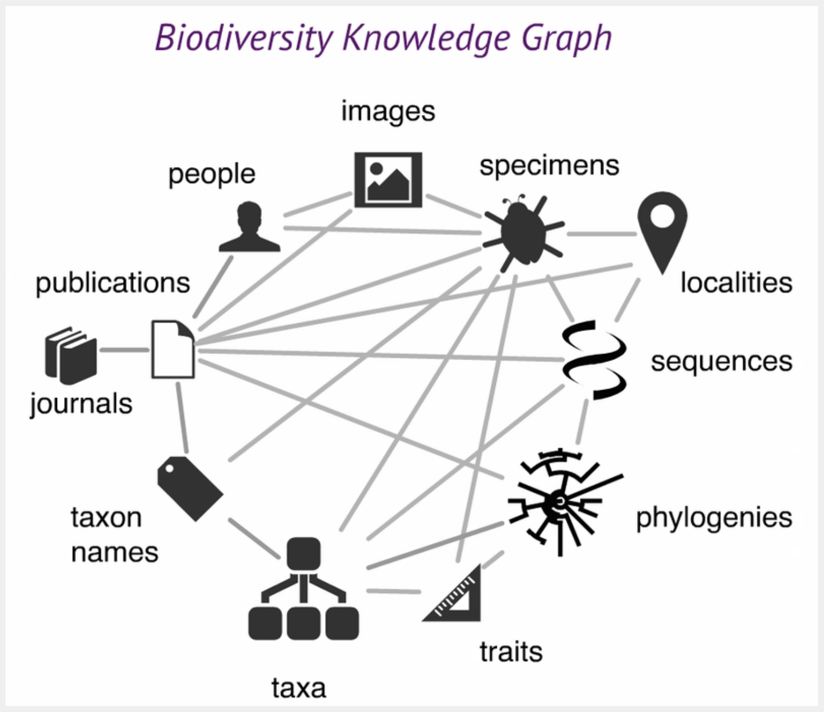


Figure: A conceptual diagram of the Biodiversity Knowledge Graph (image: Page, R. D. M. 2016)

Dr. Rod Page has advocated that we link together the diverse sources of biodiversity data into what he coins the "biodiversity knowledge graph" (Page, R. D. M. 2016) (Appendix 2: Interviews). In a recent paper, Dr. Page makes the case for Wikidata as the "logical venue for a global database of taxonomic literature, the so-called "bibliography of life."

"It not only benefits from a community of active editors, it piggy backs on the remarkable fact that taxonomy is the only discipline to have its own Wikimedia Foundation project (Wikispecies). Consequently, a large number of taxonomic works and their authors already exist in Wikidata." (Page, 2022)

To further bolster Page's argument, the number of global biodiversity databases contributing their identifiers to Wikidata is large, and growing — Wikidata's Query Service returns over 1300 biodiversity databases and catalogs interlinking their records with Wikidata – a global biodiversity hub indeed.

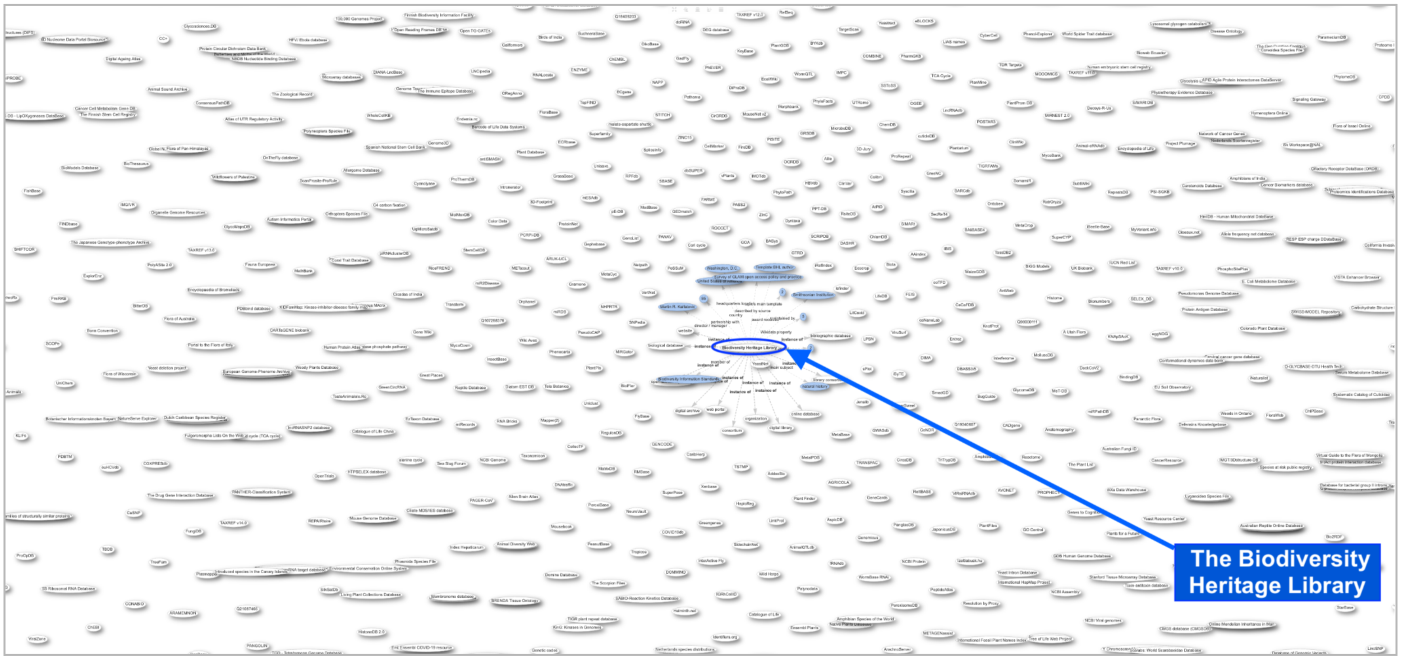


Figure: BHL amongst friends; 1300+ biodiversity databases in Wikidata. BHL is one project in the biodiversity community waiting to be further interlinked. (Image: Dearborn, 2023 https://w.wiki/6YCx)

There is much work to do to unlock, normalize, and semantically enrich the data present in BHL's corpus but, with the help of a global movement of Wikimedians, this work does not have to fall on the shoulders of "a relatively small cadre of information professionals" (Taylor et al., 2019). Demand for BHL's data is growing. Whether BHL actively participates in this work or not, it is already underway. BHL's main data outputs are already openly available. Bots and Wikimedians like user Fae, Ambrosia10, Rdmpage, Uncommon_fritillary, Magnus_Manske, and many others are bulk-loading BHL's data into the Wikimedia ecosystem.

As a major data consumer and provider, BHL's data should be measured against Berners-Lee's 5-star rating scheme, pushed into the Wikimedia ecosystem, and interwoven into the growing global web of data. Integrating better quality BHL data into Wikimedia will also fulfill BHL's commitments to the Bouchout Declaration, signed in 2014 to:

> *"promote free and open access to data and information about biodiversity by people and computers and to bring about an inclusive and shared knowledge management infrastructure that will allow our society to respond more effectively to the challenges of the present and future." (Bouchout Declaration)*

Wikimedia is a shared information infrastructure that can proliferate the Biodiversity Heritage Library's data-rich collection beyond its repository. By transforming BHL's dark data into 5-star linked open data, BHL can realize its goal to become "the most comprehensive, reliable, and reputable" resource to support global challenges.

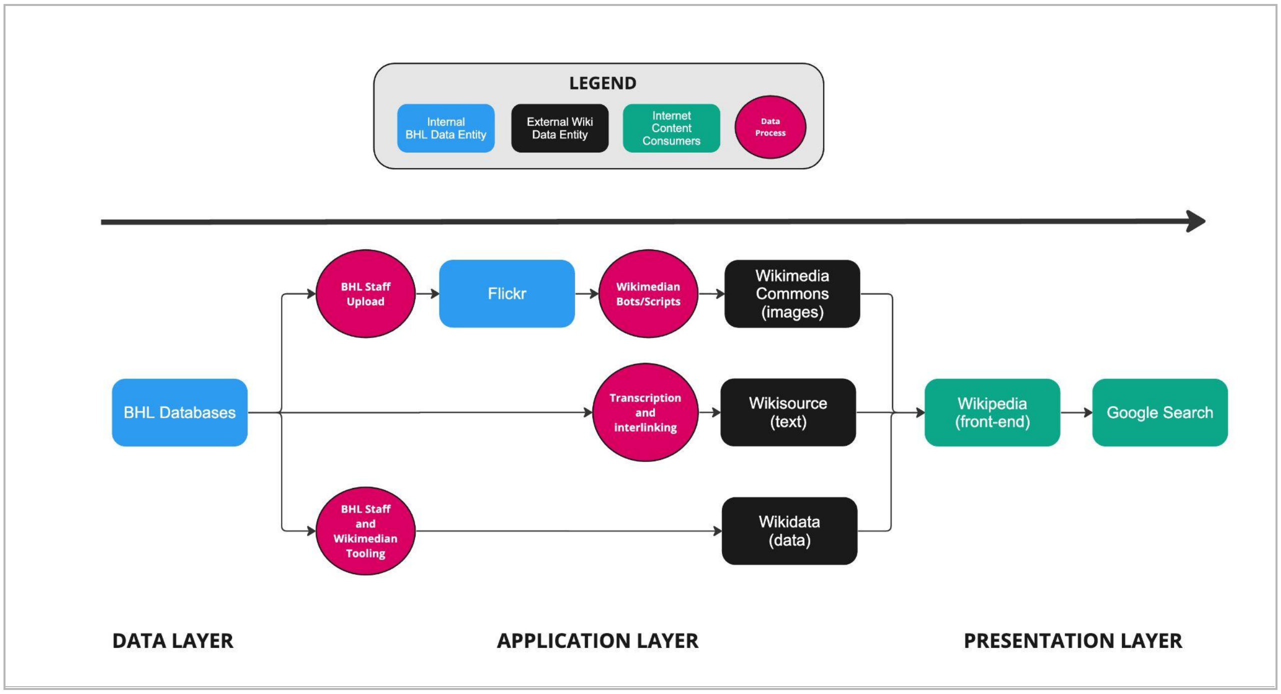


Figure: A BHL data flow diagram modeled after traditional web architecture for a single site, charting BHL's data journey to the broader web. (Image: Dearborn, 2023)

The challenges ahead will require the BHL community to act strategically and pursue new capacity-building partnerships. To overcome the digital constraints of legacy or missing metadata, prohibitive paywalls, data silos, and proprietary software and formats, BHL needs to extend beyond digitization and curation efforts within its siloed data ecosystem to bridge knowledge gaps and set biodiversity knowledge truly free.

Acknowledgments

I'm honored and indebted to the many people who took the time to read, contribute, and refine the ideas for this white paper. It was a learning journey! Thank you:

- **BHL Secretariat:** Bianca Crowley (SLA/BHL), Colleen Funkhouser(SLA/BHL), Martin Kalfatovic (SLA/BHL)

- **Mentors:** Joseph DeVeer (Harvard University, Ernst Mayr Library/BHL), Mike Lichtenberg (BHL), Joel Richard (SLA), Constance Rinaldo (Harvard University, Ernst Mayr Library/BHL), Curt Savoie (Citizen Data), Sienna Svob (Blue State), Judy Warnement (Harvard University, Botany Libraries)

- **Interviewees:** Grace Costantino (Independent), Diana Duncan (Chicago Field Museum/BHL), James Hare (Internet Archive), Siobhan Leachman (Independent Wikimedian), Andy Mabbett (Independent Wikimedian), Katie Mika (Harvard University), Dr. Rod Page (University of Glasgow), Carolyn Sheffield (The Nature Conservancy), Alex Stinson (Wikimedia Foundation), and Andra Waagmeester (Micelio/Independent Wikimedian).

- **Peer reviewers:** Bianca Crowley (SLA/BHL), Jacqueline Chapman (SLA), Sandra Fauconnier (Wikimedia Foundation), Giovanna Fontenelle (Culture and Heritage, Wikimedia Foundation), Colleen Funkhouser (SLA/BHL), Nicole Kearney (Museums Victoria/BHL Australia), Martin Kalfatovic (SLA/BHL), Siobhan Leachman (Independent Wikimedian), Mike Lichtenberg (BHL), Susan Lynch (BHL), Richard Naples, (SLA/BHL), Jake Orlowitz (WikiBlueprint), Dr. Rod Page (University of Glasgow), Suzanne Pilsk (SLA/BHL), Connie Rinaldo (BHL), Fiona Romeo (Culture and Heritage, Wikimedia Foundation), Jackie Shieh (SLA).

- **Global Names:** Dima Mozzherin (MBL) and Dr. Geoff Ower (UIUC/CoL)

- **All BHL and SLA staff,** in particular, the dedicated members of the
  - **BHL Cataloging and Metadata Committee**
  - **BHL Collections Committee**
  - **BHL Executive Committee**
  - **BHL Persistent Identifier Working Group**
  - **BHL Partners**
  - **BHL Technical Team**
  - **BHL Transcription Upload Tool Working Group**

Appendix 1: Summary of Recommendations

A list of recommendations for the BHL Consortium's consideration given its current capacities; with additional resources BHL could expand its Wikimedia efforts to meet pressing global challenges. Each of these recommendations was evaluated using the MoSCoW methodology (M - Must have, S - Should have, C - Could have, W - Won't have.)

#	Wikimedia Project	Recommendation	Priority (MoSCoW)
1	Wikipedia	Continue to hold edit-a-thons for Wikidata and Wikipedia and explore a transcribe-a-thons for Wikisource	**Could**
2	Wikipedia	Create a BHL GLAM project page	**Could**
3	Wikipedia	Solicit a Wikipedian-in-residence position	**Should**
4	Wikipedia	Add BHL to the GLAM/Smithsonian project to-do list	**Should**
5	Wikimedia Commons	Reconcile the Flickr image tag dataset to Wikimedia Commons species illustrations to further populate Wikidata and Wikipedia species pages BHL illustrations.	**Should**
6	Wikimedia Commons	Transform wikitext into structured data on Wikimedia Commons to expose BHL images to a SPARQL endpoint for user search and eventual harvest back into the BHL ecosystem.	**Should**

7	Wikidata / General	Participate in emerging linked open bibliographic data for production by electing BHL representatives to participate in the LD4P community meetings and discussions	**Should**
8	Wikidata	Review Wikidata BHL properties and make recommendations to the Wikidata community for additions to other indexed entities in BHL that do not currently have properties: Collections, Contributors, Subjects, Items	**Could**
9	Wikidata	Increase capacity towards crowdsourcing the assignment of Q#s to article metadata by documenting the workflow on the BHL Wikidata Project page and using Rod Pages BHL2Wiki tool and QuickStatements.	**Doing**
10	Wikidata	Continue to interlink BHL data growing global knowledge graph through periodic loads of updated metadata in the Mix'N'Match tool (currently: 1. BHL Bibliography IDs and 2. Creator IDs.)	**Doing**
11	Wikidata	Focus BHL efforts on the most high-impact actions in Wikidata by organizing targeted curation activities on BHL's Wikidata project page.	**Doing**
12	Wikidata	Expand capacity for current Creator ID disambiguation work (e.g. Wikidata sub-project page, BHL trainings and videos, Mix'N'Match and Open Refine workshop)	**Doing**
13	Wikidata	Rapidly develop BHL staff understanding of emerging linked open data standards by sponsoring a group session for interested BHL Staff in the Wikidata Institute	**Doing**
14	Wikidata	Harvest Wikidata QIDs and other PIDs associated with BHL creator IDs; import the matches into the BHL database	**Doing**
15	Wikidata	Connect BHL resources to other knowledge bases for our users by exposing the Wikidata QID on the BHL front-end; start the core entities: authors, articles (parts), and titles.	**Doing**

16	Wikidata	Endorse the collective goal to co-create and develop a single linked open data network for art, culture, and science by signing the WikiLibrary Manifesto.	**Should**
17	Wikidata	Expand Wikidata knowledge and efforts by soliciting a Wikidatan-in-residence position for 2022-2023, and beyond.	**Should**
18	Wikibase	Deploy a BHL Wikibase.cloud instance to explore semantic enrichment of entities not represented in the BHL data model and federate BHL's data with the growing global knowledge graph.	**Could**
19	Wikibase	Deepen understanding of BHL's relationship network by using Contributor data to construct a knowledge graph in Wikibase. Start organizations, then interlink people.	**Doing**
20	Wikibase	Understand development cost and level of effort by monitoring SLA's Minimum Viable Product (MVP) Wikibase installation; select a BHL representative to participate and observe.	**Doing**
21	General	Become eligible for extended Wikimedia benefits open to official Wikimedia User Groups and bring together Wikimedia biodiversity practitioners by applying to become a Wikimedia User Group https://meta.wikimedia.org/wiki/Wikimedia_user_groups	**Should**
22	General	Address BHL Data challenges by sponsoring annual Data Science contests; sample projects could include methods for entity disambiguation, adding geolocation data to the corpus, OCR correction, building on the art of life work, etc. (suggestion from R. Page)	**Could**
23	General	Promote and continue work on BHL's open citations and linked bibliographic data serve free knowledge by applying for a WikiCite grant	**Could**
24	General	Develop BHL staff expertise by sponsoring the acquisitions of skills and certifications focused on BHL collection and technical development from organizations like SNAC School, Scrum Alliance, Library Carpentry, Library Juice Academy, etc.	**Must**

Appendix 2: Interviews

Interviews conducted to gather key recommendations, use cases, and information about the Wikimedia ecosystem. (Ordered by date)

Person	Topic	Date
Carolyn Sheffield, Former BHL Program Manager, and Senior Information Technology Project Manager at The Nature Conservancy	*BHL User Needs Analysis*	September 9, 2021
Katie Mika, Data Services Librarian at IQSS, Harvard University	*OCR Transcription and Correction Platforms*	September 15, 2021
Siobhan Leachman, Independent Wikimedian	*Citizen Science, Notable Women, and the Wikimedia Ecosystem*	October 4, 2021
Dr. Rod Page, Professor of Taxonomy at Glasgow University and Member of BHL's Persistent Identifier Working Group	*Building the Biodiversity Knowledge Graph and Recommendations for the BHL Community*	October 13, 2021
Andy Mabbett, Independent Wikimedian	*Key Projects in the Wikimedia Ecosystem*	October 21, 2021

James Hare, Wikibase Developer / Product Management at Internet Archive	*Wikibase 101*	November 16, 2021
Grace Costantino, Former BHL Communications Manager	*BHL's Capacity for Collaboration with Wikipedia & Wikimedia Commons*	November 29, 2021
Diana Duncan, Former BHL Cataloging Chair and Chicago Field Museum Technical Services Librarian	*Overview of the BHL Author Names Project*	December 2, 2021
Andra Waagmeester, Data Scientist / Bioinformatician at Micelio and Independent Wikimedian	*Perspectives on Wikimedia Projects from a Bioinformatician*	March 23, 2022
Alex Stinson, Senior Program Strategist at Wikimedia Foundation	*Intersections: Climate Change, Open Knowledge, and Human Rights*	March 27, 2023

Appendix 3: Statistics

Statistics were gathered to understand Wikimedia project size, breadth, and active user base. To view all statistics collected please refer to the statistics workbook: *[Wikimedia Statistics as of March 2023]*

Wikidata	
Project Statistics	
Statistic	*Count*
Total # items	102,149,672
Total # of edits	1,847,574,544
Active user base	24,671
Annual page views (Feb 2022 - March 2022)	4,875,588,485
Annual unique devices (Feb 2022 - March 2022)	35,058,533

Wikimedia Commons	
Project Statistics	
Statistic	*Count*
Total # items (content pages)	89,223,876
Total # of edits	737,890,242
Active user base	40,500
Annual page views (Feb 2022 - March 2022)	12,349,946,403
Annual unique devices (Feb 2022 - March 2022)	353,696,348

Wikisource - English

Project Statistics	
Statistic	*Count*
Total # items (content pages)	988,369
Total # of edits	13,052,867
Active user base	381
Annual page views (Feb 2022 - March 2022)	203,200,488
Annual unique devices (Feb 2022 - March 2022)	12,632,940

Wikipedia - English

Project Statistics	
Statistic	*Count*
Total # items (content pages)	6,628,163
Total # of edits	1,139,032,460
Active user base	129,101
Annual page views (Feb 2022 - March 2022)	10,486,671,793
Annual unique devices (Feb 2022 - March 2022)	124,704,555,354

Wikispecies

Project Statistics	
Statistic	*Count*
Total # items (content pages)	834,984
Total # of edits	9,138,403
Active user base	200
Annual page views (Feb 2022 - March 2022)	135,098,922
Annual unique devices (Feb 2022 - March 2022)	4,809,990

Note: The Wikispecies Project was not covered as a Wikimedia "core infrastructure" project in the white paper. However, it should be noted that Wikispecies is an invaluable source of taxonomic data and work is currently underway to make the knowledge contained in the repository more actionable in the Wikimedia ecosystem and the broader web. (See Appendix 2: Interviews - Andy Mabbett)

References

About Biodiversity Heritage Library. (n.d.). https://about.biodiversitylibrary.org/

About Share-VDE. (2021). https://www.svde.org/about/about-share-vde

Alípio, S., Abdulai, M. S., Burnett, G., & Shick, D. (2021, April 14). Wikibase: the Software for Open Data projects - Wikimedia Tech News. Wikimedia Tech News. https://tech-news.wikimedia.de/en/2021/04/14/wikibase-the-software-for-open-data-projects/

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. Government Information Quarterly, 32(4), 399–418. https://doi.org/10.1016/j.giq.2015.07.006

Berners-Lee, T. [TED T. (2009). Tim Berners-Lee: The next web of open, linked data. https://www.youtube.com/watch?v=OM6XIICm_qo

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), 34–43. https://doi.org/10.1038/scientificamerican0501-34

The Biodiversity Heritage Library. (n.d.). About the Biodiversity Heritage Library. https://about.biodiversitylibrary.org/

Bohm, S. N. (2016, June 2). Why I proofread poetry at Wikisource. Diff. https://diff.wikimedia.org/2016/06/02/why-i-proofread-poetry-wikisource/

Bouchout Declaration. (2014, June 12). https://www.bouchoutdeclaration.org/

Burn-Murdoch, J. (2017). Study: Less than 1% of the world's data is analysed, over 80% is unprotected.

https://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume

Carter, P. (2021). Climate change tracking worst-case scenario. https://www.youtube.com/watch?v=fliCxyAwBWU

Chaos Computer Club. (2021). #rC3 - Wikidata for (Data) Journalists. https://www.youtube.com/watch?v=oiVOG3FuUFQ

Commons: Licensing. (n.d.). https://commons.wikimedia.org/wiki/Commons:Licensing

Commons:Pattypan - Wikimedia Commons. (n.d.). https://commons.wikimedia.org/wiki/Commons:Pattypan

Commons: project scope. Wikimedia Commons. (n.d.). Retrieved September 7, 2022, from https://commons.wikimedia.org/wiki/Commons:Project_scope#

Costantino, G. (2015, October). Tag Wikipedia – Biodiversity Heritage Library Blog. Biodiversity Heritage Library. Retrieved February 23, 2022, from https://blog.biodiversitylibrary.org/tag/wikipedia

Costantino, G. (2018, November 27). We Need Books To. . .Support Conservation. Biodiversity Heritage Library Blog. https://blog.biodiversitylibrary.org/2014/12/we-need-books-tosupport-conservation.html#more-344

Dearborn, J., & Leachman, S. (2023, March 8). The Biodiversity Heritage Library is Round Tripping Persistent Identifiers with the Wikidata Query Service. Diff. https://diff.wikimedia.org/2023/02/14/the-biodiversity-heritage-library-is-round-tripping-persistent-identifiers-with-the-wikidata-query-service/

Dearborn, J., & Kalfatovic, M. (2022, April 27). Internet Archive - Library as Laboratory Series: Analyzing Biodiversity Literature At Scale. Internet Archive. https://archive.org/details/analyzing-biodiversity-literature-at-scale

Dearborn, J., & Mika, K. (2022). Extracting expedition log data found in the Biodiversity Heritage Library. In Smithsonian Research Online. Society for the Preservation of Natural History Collections (SPNHC). https://repository.si.edu/handle/10088/114751

Diffenbaugh, N. S., & Barnes, E. A. (2023). Data-driven predictions of the time remaining until critical global warming thresholds are reached. Proceedings of the National Academy of Sciences of the United States of America, 120(6). https://doi.org/10.1073/pnas.2207183120

Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., Bermudez, L. E., & Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. Earth Science Informatics, 4(3). https://doi.org/10.1007/s12145-011-0083-6

Edmonds, H. K., Lovell, C. a. K., & Lovell, J. E. (2022). The Inequities of National Adaptation to Climate Change. Resources, 12(1). https://doi.org/10.3390/resources12010001

Freeland, C. (2012). Linking to Biodiversity Heritage Library from Wikipedia. Biodiversity Heritage Library. https://blog.biodiversitylibrary.org/2012/03/linking-to-biodiversity-heritage-library-from-wikipedia.html

Gadelha, L. M. R., Siracusa, P. C., Dalcin, E. C., Silva, L. A. E., Augusto, D. A., Krempser, E., Affe, H. M., Costa, R. L., Mondelli, M. L., Meirelles, P. M., Thompson, F., Chame, M., Ziviani, A., & Siqueira, M. F. (2020). A survey of biodiversity informatics: Concepts, practices, and challenges. WIREs Data Mining and Knowledge Discovery, 11(1). https://doi.org/10.1002/widm.1394

GBIF. (2018, July 6). Big data for biodiversity: GBIF.org surpasses 1 billion species occurrences. GBIF News. https://www.gbif.org/news/5BesWzmwqQ4U84suqWyOQy/big-data-for-biodiversity-gbiforg-surpasses-1-billion-species-occurrences

GBIF. (2019). Data from GBIF network bolsters biodiversity findings of IPCC special report. https://www.gbif.org/data-use/51171KRV0EhwmHubpfDs3h/data-from-gbif-network-bolsters-biodiversity-findings-of-ipcc-special-report

Global Names Resolver: Data Sources. (n.d.). Global Names. https://resolver.globalnames.org/data_sources

GO FAIR initiative. (2020). GO FAIR initiative: Make your data & services FAIR.

https://www.go-fair.org/

Google. (n.d.). Google Knowledge Graph Search API. Google Cloud. https://cloud.google.com/enterprise-knowledge-graph/docs/search-api

Google. (2023). How Google's Knowledge Graph works - Knowledge Panel Help. https://support.google.com/knowledgepanel/answer/9787176?hl=en#:~:text=In%20addition%20to%20public%20sources,knowledge%20panels%20they've%20claimed.

Google. (n.d.). *Margaret Mead Knowledge Panel*. https://www.google.com/search?q=margaret+mead&rlz=1C5GCEM_enUS1004US1004&oq=margaret+mead&aqs=chrome.0.0i271j46i433i512j35i39j0i433i512j0i512l2j46i340i512l2j0i512l2.3284j0j7&sourceid=chrome&ie=UTF-8

Groom, Q., Güntsch, A., Huybrechts, P., Kearney, N., Leachman, S., Nicolson, N., Page, R. D. M., Shorthouse, D. P., Thessen, A. E., & Haston, E. (2020). People are essential to linking biodiversity data. Database, 2020. https://doi.org/10.1093/database/baaa072

Haller, A., Polleres, A., Dobriy, D., Ferranti, N., & Mendez, S. (2022). An Analysis of Links in Wikidata [Paper]. European Semantic Web Conference. https://2022.eswc-conferences.org/wp-content/uploads/2022/05/paper_41_Haller_et_al.pdf

Harrisson, T. (2021, October 11). Explainer: The high-emissions 'RCP8.5' global warming scenario. Carbon Brief. https://www.carbonbrief.org/explainer-the-high-emissions-rcp8-5-global-warming-scenario/

Hausenblas, M. (2012). 5-star Open Data. https://5stardata.info/en/

International Data Corporation. (2018). Digitization of the World | From Edge to Core [Techreport]. International Data Corporation. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-white paper.pdf

IPCC report: 'Code red' for human-driven global heating, warns UN chief. (2021, December 14). UN News. https://news.un.org/en/story/2021/08/1097362

Jackson, E., & Costantino, G. (2019, March 18). #HerNaturalHistory: Open Data, BHL, and Wiki Projects. Biodiversity Heritage Library. https://blog.biodiversitylibrary.org/2019/03/hernaturalhistory-wiki-projects.html#more-21965

Johnson, I., Perry, N., Gordon, K., & Katz, J. (2021, September 22). Searching for Wikipedia: DuckDuckGo and the Wikimedia Foundation share new research on how people use search engines to get to Wikipedia. Diff. https://diff.wikimedia.org/2021/09/23/searching-for-wikipedia-duckduckgo-and-the-wikimedia-foundation-share-new-research-on-how-people-use-search-engines-to-get-to-wikipedia/

Johnson, K. R., & Owens, I. F. P. (2023). A global approach for natural history museum collections. Science, 379(6638), 1192–1194. https://doi.org/10.1126/science.adf6434

Kearney, N., Funkhouser, C., Lichtenberg, M., Missell, B., Page, R. D. M., Richard, J., Rielinger, D., & Lynch, S. V. (2021). #RetroPIDs: The missing link to the foundation of biodiversity knowledge. Biodiversity Information Science and Standards, 5. https://doi.org/10.3897/biss.5.74141

Keeler, K. (2023, February 2). How Wikipedia Distorts Indigenous History. Slate Magazine. https://slate.com/technology/2023/02/wikipedia-native-american-history-settler-colonialism.html

Keyes, O. (2019, October 26). Keynote: Questioning Wikidata. https://media.ccc.de/v/wikidatacon2019-15-keynote_questioning_wikidata

LD4P. (2020). LD4P3 Sinopia Project Plan. https://docs.google.com/presentation/d/1GFNXBZj-YTyxr8An-590kbieHzI8jvPoZp-I_gBR_iY/edit#slide=id.g8697bb5346_0_546

LD4P Partners. (2016). Linked Data for Production (LD4P) - LD4P public website - LYRASIS Wiki. https://wiki.lyrasis.org/display/LD4P

Leachman, S. (2018). How A Citizen Scientist Can Reuse & Link Biodiversity Heritage Library Data. Biodiversity Information Science and Standards. https://doi.org/10.3897/biss.2.25298

Leachman, S. (2019, October). HerNaturalHistory - Citizen Scientist Siobhan Leachman on how Wikidata can play a vital role in surfacing notable women [Video]. The University of Edinburgh. https://media.ed.ac.uk/media/HerNaturalHistory+-+Citizen+Scientist+Siobhan+Leachman+on+how+Wikidata+can+play+a+vital+role+in+surfacing+notable+women/1_pjulqj94

Library of Congress. (n.d.). New BIBFRAME-to-MARC Conversion Tools. https://www.loc.gov/bibframe/news/bibframe-to-marc-conversion.html

Marshall, M. (2021). Women in Historical SciArt: BHL Empowers Research on Women in Scientific Illustration. https://blog.biodiversitylibrary.org/2021/03/women-in-histsciart.html#more-28153

McClanahan, P. (2018). Final Report: BHL User Needs Analysis [Techreport]. The Biodiversity Heritage Library. https://drive.google.com/drive/folders/1bFiwmRiU-Jy-arStUhr0MYzD-yn5OLtZ

Miller, E., Ogbuji, U., Mueller, V., & MacDougall, K. (2012). Bibliographic framework as a web of data: Linked data model and supporting services [Techreport]. Library of Congress. https://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

Mulligan, J., Ellison, G., Levin, K., Lebling, K., Rudee, A., & Leslie-Bole, H. (2023, March 17). 6 Ways to Remove Carbon Pollution from the Atmosphere. World Resources Institute. https://www.wri.org/insights/6-ways-remove-carbon-pollution-sky

Nemet, G., Deich, N., Cohen-Brown, N., Anderson, A., & World Resources Institute. (2023, March). Carbon Removal at Scale: A Call to Action from the IPCC Report. In Panel Discussion.

O'Neil, M. (2021, November). Students are told not to use Wikipedia for research. But it's a trustworthy source. The Conversation. https://theconversation.com/students-are-told-not-to-use-wikipedia-for-research-but-its-a-trustworthy-source-168834

Page, R. D. M. (2009, December 21). iPhylo: BioStor. https://iphylo.blogspot.com/2009/12/biostor.html

Page, R. D. M. (2011). Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. BMC Bioinformatics, 12(1).

https://doi.org/10.1186/1471-2105-12-187

Page, R. D. M. (2016). Towards a biodiversity knowledge graph. Research Ideas and Outcomes, 2, e8767. https://doi.org/10.3897/rio.2.e8767

Page, R. D. M. (2022). Wikidata and the bibliography of life. PeerJ, 10, e13712. https://doi.org/10.7717/peerj.13712

Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., & Pintscher, L. (2016). From Freebase to Wikidata. Proceedings of the 25th International Conference on World Wide Web. https://doi.org/10.1145/2872427.2874809

Perez, S. (2012). Wikipedia's Next Big Thing: Wikidata, A Machine-Readable, User-Editable Database Funded By Google, Paul Allen And Others. https://techcrunch.com/2012/03/30/wikipedias-next-big-thing-wikidata-a-machine-readable-user-editable-database-funded-by-google-paul-allen-and-others/

Peterson, A. T., Soberón, J., & Krishtalka, L. (2015). A global perspective on decadal challenges and priorities in biodiversity informatics. BMC Ecology, 15(1). https://doi.org/10.1186/s12898-015-0046-8

Pintscher, L., Voget, L., Koeppen, M., & Aleynikova, E. (2019). Strategy for the Wikibase Ecosystem. In Wikimedia. https://upload.wikimedia.org/wikipedia/commons/c/cc/Strategy_for_Wikibase_Ecosystem.pdf

Pratt University, Semantic Lab. (n.d.). Linked Jazz | Revealing the Relationships of the Jazz. Community. https://linkedjazz.org/

Redi, M. (2018, August 17). How we're using machine learning to visually enrich Wikidata. Wikimedia Foundation. https://wikimediafoundation.org/news/2018/03/14/machine-learning-visually-enriching-wikidata/

Richard, J. (2022, December 20). OCR Improvements: An Early Analysis. Biodiversity Heritage Library. https://blog.biodiversitylibrary.org/2022/07/ocr-improvements-early-analysis.html

Rives, K. (n.d.). World must cut 28 gigatons of carbon in 8 years to meet climate goal, UN says. S&P Global Market Intelligence. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/world-must-cut-28-gigatons-of-carbon-in-8-years-to-meet-climate-goal-un-says-67252932

Schroeer, T. (2017, September 8). Cognitive computing: Hello Watson on the shop floor. Retrieved September 25, 2021

Schreur, P. (2018, July). The Evolution of BIBFRAME: from MARC Surrogate to Web Conformant Data Model . http://library.ifla.org/id/eprint/2202/1/141-schreur-en.pdf

Sengel-Jones, M. (2021). The promise of Wikidata: How journalists can use the crowdsourced database. https://datajournalism.com/read/longreads/the-promise-of-wikidata

Shieh, J. S. (2022, September 14). Library Descriptive Data Unleashed in SVDE: Making Data More Meaningful – Smithsonian Libraries and Archives / Unbound. Unbound | Smithsonian Libraries and Archives. https://blog.library.si.edu/blog/2022/09/14/library-descriptive-data-unleashed-in-svde-making-data-more-meaningful/

Similarweb. (2023, January). Top Search Engine Market Share in February 2023 | Similarweb. https://www.similarweb.com/engines/

Sloan Grant. (2016). Wikimedia Commons. Retrieved March 7, 2023, from https://commons.wikimedia.org/wiki/Commons:Structured_data/Sloan_Grant

Stanford University Libraries. (2018). Sinopia Demos. https://www.youtube.com/playlist?list=PLrOZtzLTYPJcBux29ezxZju9FgcifJ2pn

The State of Carbon Dioxide Removal Report. (2023). The State of Carbon Dioxide Removal. https://www.stateofcdr.org/home/#key

Stuart, S. (2021, June). Wikipedia: The Most Reliable Source on the Internet? PC MAG. https://www.pcmag.com/news/wikipedia-the-most-reliable-source-on-the-internet

Taylor, M., Wilson, K., Hammer, S., & Gorrell, M. (2019). Proposed Revisions to the BHL Data Model [Techreport]. Index Data. https://docs.google.com/document/d/1O-I5o0uiYRvuWEpDG5O7U-kuxC9TOs_drE-YzVO kp1o/edit#heading=h.fufy0nd7qey7

Tennant, R. (2021). MARC Must Die. https://www.libraryjournal.com/?detailStory=marc-must-die

Trice, A. (2016). The Future of Cognitive Computing. https://www.ibm.com/blogs/cloud-archive/2015/11/future-of-cognitive-computing/

UCSB [Research Data Services]. (2020). Persistent Identifiers Explained. library.ucsb.edu/. https://www.library.ucsb.edu/sites/default/files/dls_n4_pids_navy.pdf

United Nations. (2021). IPCC report: 'Code red' for human-driven global heating, warns UN. https://news.un.org/en/story/2021/08/1097362

United Nations. (2023, March 20). SG message for the launch of the Synthesis Report of the Intergovernmental Panel on Climate Change [Video]. YouTube. https://www.youtube.com/watch?v=p_rPNTWwuLU

Vershbow, B. (2018, October 29). How could Wikimedia Commons be improved? A conversation with designer George Oates. Diff. https://diff.wikimedia.org/2018/10/29/george-oates-conversation/

Vincent, N., & Hecht, B. (2021). A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results. Proceedings of the ACM on Human-Computer Interaction, 9. https://doi.org/10.1145/3449078

W3C. (2013, March 21). SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/

W3C. (2014, February 25). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. https://www.w3.org/TR/rdf11-concepts/

WDQS Search Team. (2022). WDQS Backend Alternatives: The Process, Details, and Results. In Wikidata (Version 1.1). https://upload.wikimedia.org/wikipedia/commons/e/ea/WDQS_Backend_Alternatives_working_paper.pdf

Wikibase. (2023, April 2). MediaWiki. https://www.mediawiki.org/wiki/Wikibase

Wikidata. (2017). Wikidata:Database_download. https://www.wikidata.org/wiki/Wikidata:Database_download

Wikimedia Deutschland. (2019). Wikimedia Germany Annual Report 2019 [Techreport]. Wikimedia Deutschland. https://www.wikimedia.de/2019/en/themen/wikidata/

Wikimedia Foundation. (2012). Wikidata:Statistics - Wikidata. https://www.wikidata.org/wiki/Wikidata:Statistics

Wikimedia Foundation. (2011, December 12). The Impact of Wikipedia - Andrea Zanni [Video]. YouTube. https://www.youtube.com/watch?v=8Z9IcBmrmeY

Wikimedia Foundation. (2018). Wikimedia Foundation Mission. https://wikimediafoundation.org/about/mission/

Wikimedia Foundation. (2022, May). Category: All extensions. MediaWiki. https://www.mediawiki.org/wiki/Category:All_extensions

Wikimedia Foundation. (2022, December 27). Wikibase/Federation. MediaWiki. https://www.mediawiki.org/wiki/Wikibase/Federation

Wikipedia contributors. (2021). MediaWiki. https://en.wikipedia.org/wiki/MediaWiki

Wikipedia contributors. (2022, January 30). Semantic triple. Wikipedia. https://en.wikipedia.org/wiki/Semantic_triple

Wikipedia contributors. (2023b, April 6). Disinformation attack. Wikipedia. https://en.wikipedia.org/wiki/Disinformation_attack

Wikimedia Commons contributors. (2023, March 19). Wikimedia Commons. Wikipedia. https://en.wikipedia.org/wiki/Wikimedia_Commons

Wilkinson, M. (2016). The FAIR guiding principles for scientific data... https://www.nature.com/articles/sdata201618

White House. (2022). White House Releases First-of-a-Kind Indigenous Knowledge Guidance for Federal Agencies. The White House. https://www.whitehouse.gov/ceq/news-updates/2022/12/01/white-house-releases-first-of-a-kind-indigenous-knowledge-guidance-for-federal-agencies/

World's Most Vulnerable Nations Suffer Disproportionately. (n.d.). United Nations News. https://www.un.org/ohrlls/news/frontline-climate-crisis-worlds-most-vulnerable-nations-suffer-disproportionately