Workshop #2: Transforming Biodiversity Heritage Library Images with Structured Data on Commons (SDC)

- **Date/Time:** Wednesday, **APRIL 10 @ 2:00 PM EST |** Find your time
- **Event Page:**
  - 📄 Transforming Biodiversity Heritage Library Images - Data Modeling with O…
- **Join Zoom Meeting:**
  https://smithsonian.zoom.us/j/86563563251?pwd=NEU3MWovMkRaQW5hV09BdTlRZ1ZVdz09
- **Recording:**
  https://smithsonian.zoom.us/rec/share/pd4QonjEJvIfeOgD6c-cdGyxjq-JOfGUm1fsXFGfc_z-Pgafn-gJbqLq8THyHLB6.pJPkGnkEm7kjrzN2?startTime=1712771704000
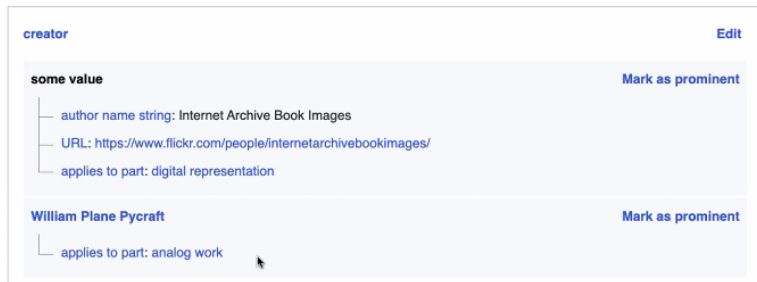 - **Passcode:** 4%h^5K@a

---

Agenda

Attendees:

- **Siobhan Leachman (Wikimedian)**
- **Michelle Marshall (@mlmarshall13 (Flickr) and @HistSciArt)**
- **JJ Dearborn (BHL)**
- **Sandra Fauconnier (Wikimedian)**
- **Greta Suiter (Ohio University)**
- **Sabine von Mering (MfN)**
- **Christine Fernsebner Eslao (Harvard)**
- **Alex Chan (Flickr Foundation)**
- **Susan Lynch (BHL and AMNH)**
- **Suzanne Pilsk (SLA)**
- **Kendra Meyer (American Museum of Natural History)**
- **Joe deVeer (Harvard/MCZ)**
- **Diane Shaw (Smithsonian retiree)**

- **Giovanna Fontenelle (Culture and Heritage, Wikimedia Foundation)**
- **Andy Mabbett, User:Pigsonthewing (Wikimedian)**
- **Nat Hernández Clavijo, User:Piracalamina (Wikimedistas de Uruguay)**
- **Kelli Trei (University of Illinois Urbana, Champaign)**

Time	Agenda Item	Link	Notes/Details
5-10	**Overview (JD/SF/GF)**		## Project Summary: BHL2SDC The goal of these workshops is to bring BHL Staff, Wikimedians, Flickr staff, and all biodiversity image enthusiasts together to begin the process of mapping and loading BHL Flickr image data to Structured Data on Commons (SDC), a new Wikimedia Commons initiative. This project was the second top-voted priority from the BHL _Wikimedia Whitepaper entitled Unifying Biodiversity Knowledge to Support Life on a Sustainable Planet_. By completing our mapping work as a community and converting wiki text to structured data, we will unlock the potential to describe BHL images more robustly while building a viable SPARQL endpoint for BHL's Image Collection. This work will improve the data ecosystems for Wikimedia Commons, Flickr, and BHL alike and most importantly serve our global user bases through enhanced image search precision and retrieval. BHL Flickr Data - https://doi.org/10.25573/data.24476686.v3
15	**Community Documentation - DRAFT - Data Model Proposal (Sandra)**	📄 BHL illustr…	**Question for Discussion** 1. How to deal with data drift between illustration records on Wikimedia Commons and Publication records in Wikidata? 2. What data will be held on the bhl illustration record and what data will be held on the wikidata publication record 3. Do we want an approach where there is no need to have corresponding Wikidata items for the publications? 4. For publications, how to deal with FRBR levels and relationship between higher-level works, volumes? 5. Do we want to consciously populate media with the full pdfs of the books? NOTES: - This is a living document that will continue on with the WiR

- 3 columns have been added to the spreadsheet
    - Recommendations
    - Notes
    - Sandra Notes
- See Siobhan's example
  https://commons.wikimedia.org/wiki/File:Archives_entomologiques,_ou,_Recueil_contenant_des_illustrations_d'insectes_nouveaux_ou_rares_(Frontispiece)_(8188502006).jpg
- There are many images in Wikimedia that are from BHL, Flickr, and IA book images (primary sources)
- Categories are used to compile images into a collection derived from a publication example:
  https://commons.wikimedia.org/wiki/Category:Archives_entomologiques,_ou,_Recueil_contenant_des_illustrations_d%27insectes_nouveaux_ou_rares
- Sandra showing the broader data ecosystem and how one image is connected to other parts of wiki
- Wikitext is "old style"; this is what SDC is rolled out to convert that data over in the structured data tab.
- From Andy: Illustrator ID is crucial; their death date determines the copyright status of the work in many countries
- Sandra looked at backfillrbot and modelled the data in the same consistent way.
- Dealing with "creators"; perhaps backfillr bot could add qualifier **"applies to part: digital representation."**

Siobhan: not loving Internet Archive as "creator" would rather have it listed with "source of file"

Siobhan: It is really important to put the data model to work in practice. Everyone pick an image and try to add some statements

JJ: See **sample data** if you want to start playing with the structured data:
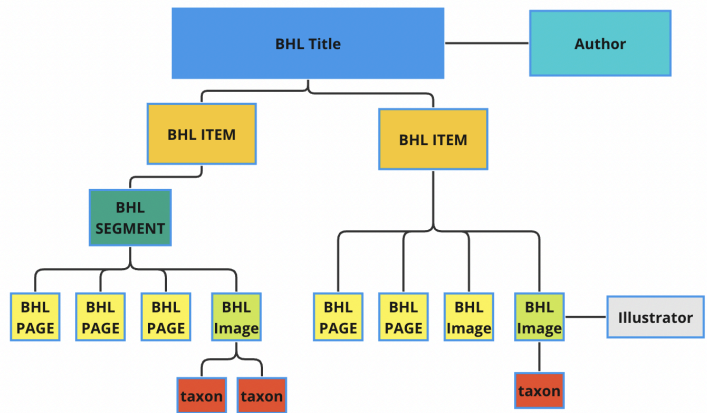📊 DRAFT - Mapping - BHL Flickr Images to Wiki SDC

| 5-10 | **Understanding Backfillrbot** | https://commons.wikimedia.org/wiki/User:FlickypediaBackfillrBot | **Questions for Alex:**

 ● **How does the bot know which images to work on?**

 Wikimedia Commons publishes a snapshot of the entire dataset. (32GBs) They work through the whole list and they figure out if it is from FLICKR. Looking for flickr id, source id, etc. crunch everything down and find the subset that is FLICKR derived.

 ○ **Is the bot working at random through images?**

 The bot works through the sequence (numeric id) roughly corresponds to oldest to newest based on upload date.

 ○ **How long will it take to run through all 12M files?**

 A long time. Months, possibly a year or more.

 ● And the bot adds these properties to SDC if they are missing?
   ○ Flickr photo ID (P12120)
   ○ creator (P170)
   ○ source of file (P7482)
   ○ copyright license (P275)
   ○ copyright status (P6216)

 ● Backfillrbot is a project from last year in tandem with Flickypedia, the spiritual successor of flickr2commons.
 ● See Flickypedia data model: https://commons.wikimedia.org/wiki/Commons:Flickypedia/Data_Modeling
 ● **Backfillrbot addresses existing images on Wikimedia Commons.**
 ● This wasn't done completely done out of the goodness of our hearts, there are reasons we built.
 ● Selfish reason why we built it: Figuring out what is from FLICKR, isolating the dataset.
 ● Alex needs to do maintenance to get it unpaused. Technical issues link
 ● Siobhan: Copyright and is bot purely additive?
 ● Alex: yes it will add an additional license on top of the old one (that may have been wrong); the Bot isn't updating old licenses
 ● Siobhan: keen on license guidance (especially when it changes) - deprecate? Delete? Old incorrect metadata about copyright licenses. PDLA - modeling see below. Referencing may be a way to solve this issue.

 See: https://commons.wikimedia.org/wiki/Commons:Digital_Public_Library_of_America/Modeling |
| 30 | **Mapping remaining BHL Data to Structured Data on Commons** | 📊 DRAFT - M… | See spreadsheet NOTE column for remaining DECISIONS to be made by the community. |

	Properties (JJ/Sandra)		Schema questions - Wikimedians: Explain - captions vs. depicts vs. description? - How do we indicate page types?
10-15	**Flickypedia Demo**	https://www.flickr.org/tools/flickypedia/get_photos	Example of a file loaded with the Flickypedia tool https://commons.wikimedia.org/wiki/File:A_monograph_of_Australian_land_shells_1034543.jpg Flyckypedia requests: - **Is it possible to bring in machine tags as structured data too?** Feature request from the community RE: **Machine tags - can we at least map the machine ones to SDC?** - **Can you talk about duplicate detection?** The tool does detect duplicates and will show you the album or files if they are already on commons. DRAFT documentation 📄 Adding BHL Flickr Images to Wikimedia Commons Using Flickypedia Question: What should we use as the **Title** which becomes the **File name?** **Andy/Siobhan:** Needs to be a unique string because this maps to file name. Typically use the publication name + a plate or page number. For plate numbers, use numeric instead of roman numerals and add a leading 0 if you want them to sort properly in a large category set. With the flickypedia tool you may not want to add a url string to the description. It throws back an error. Worth further investigation.
5	**Understanding Templates: The BHL Commons Template (SF)**	https://commons.wikimedia.org/wiki/Template:BHL	Are templates still useful and used in the new SDC paradigm? Sandra:

			Ideally, such templates become minimal and fully structured-data-driven. This has not yet been developed for the biodiversity field, though. I would like to strongly argue to work on one, or just a few, generic biodiversity related template(s) rather than have many specialized ones - and thus also to not have a {{BHL}} template but a broader one.
Save for BHL-WIKI	**Dealing with duplicates**	https://commons.wikimedia.org/w/index.php?search=biodiversitylibrary.org%2Fpage%2F37141735&title=Special:MediaSearch&type=image	• What is the best course of action here? • Which one receives SDC treatment?
FYI	**Other SDC tools and bots**	https://commons.wikimedia.org/wiki/User:SchlurcherBot https://commons.wikimedia.org/wiki/User:BotMultichillT	(SF) Others that are handy for easier data entry: • AC/DC • The SDC gadget
Save for BHL-WIKI	**Connecting SDC Images to entities derived from publications in Wikidata (Everyone to discuss)**		Group Discussion: • Right now, what tasks / actionable steps happen when an image comes from a book? • What wikidata work and other named entity connections need to be made? • What other data models do we need to document for the community? And where should these be documented?

Data Entity Relationships in BHL

Workshop Prep work

- ☑ ~~sandra.fauconnier@gmail.com~~ ~~to DRAFT Data model proposal / big picture overview that will serve as an ongoing discussion doc~~
- ☑ ~~gfontenelle@wikimedia.org~~ ~~to check in with Flickr folks on their bot timeline & need for feedback from this group~~
  - ~~[Giovanna]: Data modeling:~~ ~~https://commons.wikimedia.org/wiki/Commons:Flickypedia/Data_Modeling~~
- ☑ ~~sandra.fauconnier@gmail.com~~ ~~to send to relevant telegram channels (jj not on biodiversity channel)~~
- ☑ ~~Giovanna Fontenelle~~ ~~to send to relevant technical users (interested in LUA), Flickr Foundation (and someone else?)~~
- ☑ ~~JJ Dearborn~~ ~~to add property proposal suggestion/discussion items for items in the DECIDE state.   IN PROGRESS~~

☑ ~~JJ Dearborn Send out meet-up reminder APRIL 9~~
☑ ~~JJ Dearborn to create new column in mapping spreadsheet and figure out what data is already on Commons as wikitext~~
☑ ~~JJ Dearborn to create new agenda for pt. 2~~
☑ ~~dearbornjj@si.edu Ask Flickr team if they are willing to participate in Flickypedia demo~~ ~~https://www.flickr.org/tools/flickypedia/get_photos~~ & Bot: https://commons.wikimedia.org/wiki/User:FlickypediaBackfillrBot
☑ ~~JJ Dearborn to update invite HOLD the week of MAR 25 with all pertinent details and links and invitee list (everyone from last time + BHL-WIKI); send to BHL-ALL STAFF listserv~~
☑ ~~@Everyone to review Siobhan's~~ ~~research expedition page~~ ~~and~~ ~~discussion doc~~ ~~to see if we want to create something similar for BHL illustrations on Commons~~

Workshop #1: Transforming Biodiversity Heritage Library Images - Data Modeling with OpenRefine

- Date/Time: Friday, October 27 @ 2:00 PM EST | Find your time
- Event Page: https://meta.wikimedia.org/wiki/Image_Description_Month
- Join Zoom Meeting: https://smithsonian.zoom.us/j/81157161366?pwd=QklxQW5KUUczZG16NnJNb1k3V1crdz09
- **Recording:** https://smithsonian.zoom.us/rec/share/FSKcooIdHOXjxX04-dWfE9ubgNGD4ytIv_wJNIDnDFrNuy9GryCM2cXDYkGZtCLh.HyhnU8-xi5R6s6BQ
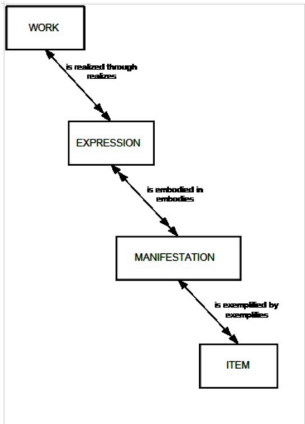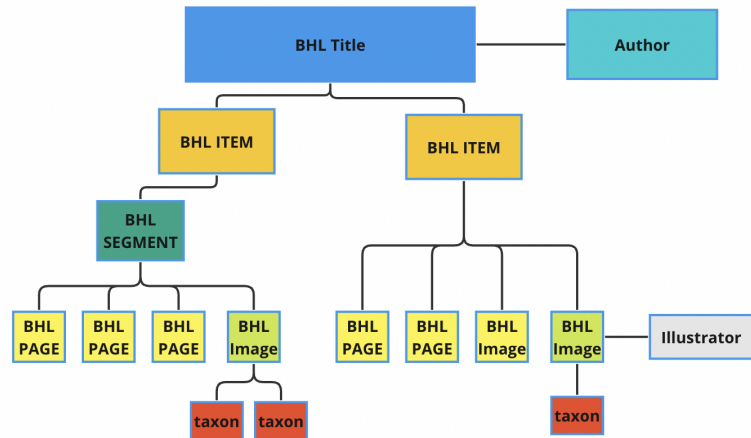 - **Passcode: !m1g$&sj**

Agenda

Attendees:

- Tamsin Braisher (User:DrThneed) braishert@yahoo.co.nz
- Siobhan Leachman (Ambrosia10) siobhan_leachman@yahoo.co.nz
- Althea Meer, NYBG
- Brianna Giasullo, ANSP
- Mike Dickison (User:Giantflightlessbirds)
- Sandra Fauconnier (User:Spinster)
- Alicia Fagerving, Wikimedia Sverige
- Jessamyn West, Flickr Foundation (User: Jessamyn)
- Suzanne Pilsk, Smithsonian (User:Pilsks)
- Susan Lynch, BHL (User:LikesLibraries)
- Richard Naples, Smithsonian (User:Drastician)
- Robert Fernandez, Wikimedia District of Columbia (User:Gamaliel) wikigamaliel@gmail.com
- Daniel Euphrat, SLA
- Giovanna Fontenelle, WMF
- Brodie Hoare, Auckland Museum
- JJ Dearborn, BHL (User:JJFord BHL)
- Diane Shaw (User Uncommon_fritillary; retired Smithsonian librarian)

Time	Who	Agenda Item	Notes/Details
10	Giovanna & Sandra	About Structured Data on Commons	The who, what, where, when, why + progress so far on image modelling Structure Data on Commons: About https://commons.wikimedia.org/wiki/Commons:Structured_data We can describe files now with structured linked data similar to Wikidata now (essentially a Wikibase can now store structured data around Images) It is not exactly the same as Wikidata. Often people ask if we should create an entity in Wikidata for images but the answer is NO. Wikidata and SDC are linked but not the same. Any questions?

			● Giovanna - Started as project back in 2017; 2020 - Current - Project "Structured Data Across Wikimedia." ● Trying to revamp the search on Wiki Commons also show images from Commons in Wikidata.
10	JD	BHL Data Exports of Image Data	Link to BHL Flickr Data - https://doi.org/10.25573/data.24476686.v3 Combined, the files are about 65MB zipped, and still 1.4GB expanded. Each individual file, when expanded, is about 350MB. There are 45 fields to be mapped to SDC. We pulled data from the Title, Item, Page, Names tables and combined them with the Flickr harvest table on our dev server. It's a lot of data and chunked out into 4 files so that we can work them in OpenRefine.
20	JD	Data Modelling considerations for images derived from publications	Consider: LRM IFLA https://www.isko.org/cyclo/lrm and https://www.wikidata.org/wiki/Wikidata:WikiProject_Books Work/ Expression - PSA - these levels of data can/should be separate entities in Wikidata See: Cataloging Scenarios in BHL to model in Wikidata And

Data Entity Relationships in BHL

JD - Scientific names are created using Global Names Algorithms/ APIs and stored in the BHL Database along side the Page. We have name, page ID pairs in the dataset above.

DE - The sci name data in Flickr is crowdsourced using the machine tags

JD - We store both flavors of names - we have Global Names data + the Flickr machine taxotags now in those data dumps (above).

Siobhan L - Machine tags are broken in FLickr and won't index in Flickr any more

Siobhan L - Many of the tags are synonyms because BHL is historic literature and the authorized name is not necessarily there. **In a perfect world we want both the synonym + the canonical authorized form of the name.**

Siobhan L - Soapbox Often these pubs have a lot of author data but not illustrator data.

			*Suggestion: BHL images should share the DOI of their publication? Plate should have a PID Sandra F - Today is scratching the surface on these data modelling conversations. Let's also flag major questions that we can answer later? **How important is it to have full BHL data represented in SDC?** JD - Preliminary thoughts are that it is important to store the IDS for sure because they are gateways to more data if needed; the rest of the data that we bring over to SDC is really about user search and retrieval. What is going to help our users find what they are looking for?

45	JD	Mapping BHL Data to Structured Data on Commons Properties	See: [BHL Data Mapping to SDC Properties Workbook](#)
			
			Handy tool - to browse for properties use: https://hay.toolforge.org/propbrowse/
			
			Reference: What other image schema out there in the wild?
			
			- [IPTC Metadata](#)
			- [DC Metadata](#)
			- [Schema.org](#)
			- [Wikidata Cradle Template: Artwork](#)
			- [DPLA Data Modelling](#)
			- Structured Data on Commons data modeling conventions:
			    ○ General https://commons.wikimedia.org/wiki/Commons:Structured_data/Modeling
			    ○ The basics for every file https://commons.wikimedia.org/wiki/Commons:Structured_data/Modeling#The_basics:_structured_data_for_every_Commons_file
			    ○ Uploads from Flickr https://commons.wikimedia.org/wiki/Commons:Structured_data/Modeling#Upload_from_a_platform_like_Panoramio,_Geograph_or_Flickr
			    ○ Data model by the DPLA https://commons.wikimedia.org/wiki/Commons:Digital_Public_Library_of_America/Modeling
			
			ACTION: Check-in with Dominic and Bianca around Copyright/Right Statements for BHL to SDC mapping.

Next time	JD/RN	Examples provided by Jackie Shieh of entity relationships around images derived from publications	**1. Image: BHL Flickr Image** https://www.biodiversitylibrary.org/page/43395695#page/307/mode/1up **2. Image: Wikimedia Commons** https://commons.wikimedia.org/wiki/File:Mary_Vaux_Walcott_-_Spotted_Bee_Balm_(Monarda_punctata)_-_1970.355.571_-_Smithsonian_American_Art_Museum.jpg **3. Work in Wikidata** https://www.wikidata.org/wiki/Q42193529 **4. Publication in Wikidata** https://www.wikidata.org/wiki/Q51511013 **5. Illustrator: Mary Vaux Walcott** https://www.wikidata.org/wiki/Q6780882 https://www.wikidata.org/wiki/Q1760733 **6. Common Name - Spotted Bee Balm (from SAAM watercolor)** https://www.wikidata.org/wiki/Q20518257 Q: Which entities are important to connect? What other entities could be interlinked?
Next time	SF	Images from Publications in Wikidata	Right now, what tasks / actionable steps happen when an image comes from a book? What wikidata work and connections need to be made?
20	SF/JD	OpenRefine Demo w/BHL Flickr data	Download Open Refine here: https://openrefine.org/ Commons extension here: https://github.com/OpenRefine/CommonsExtension Examples shown by Sandra: ● Publication with illustrations on BHL: ○ Work level Wikidata item https://www.wikidata.org/wiki/Q123196475 - Commons category https://commons.wikimedia.org/wiki/Category:Familiar_trees_and_their_leaves

			○ Edition level Wikidata item (should we go more in detail than this / lower in FRBR ranks)? https://www.wikidata.org/wiki/Q123207411 - Commons category https://commons.wikimedia.org/wiki/Category:Familiar_trees_and_their_leaves_(1911) ○ Other publication demo-ed in OpenRefine, edition level item - Commons category https://commons.wikimedia.org/wiki/Category:Collectio_plantarum_tam_exoticarum,_quam_indigenarum,_cum_delineatione,_descriptione_culturaque_earum ● One Wikimedia Commons file with some manually added structured data (not according to shared modeling consensus yet) https://commons.wikimedia.org/wiki/File:Familiar_trees_and_their_leaves,_described_and_illustrated_by_F._Schuyler_Mathews,_with_illus._in_colors_and_over_two_hundred_drawings_by_the_author,_and_an_introd._by_L.H._Bailey_(Page_10)_(6254946484).jpg SL - PSA the artist is not the author of the publication title.