Australasian Open Access Strategy Group - Open Access Week 2020
Open with Purpose: Taking Action to Build Structural Equity and Inclusion

Open Access and Biodiversity knowledge by Siobhan Leachman
https://orcid.org/0000-0002-5398-7721

[Slide 1]
I'm Siobhan Leachman - a Wikimedian and citizen scientist here to talk to you about HOW Open Access benefits the work I do with biodiversity information and data.

Many folk believe Open Access means access to scientific literature. And it does, but I want to discuss how Open Access is SO much more generous than that.

[Slide 2]
This is how Open Access is defined in Wikipedia. I take a very wide view of what constitutes a "research output". I want access to not just scholarly articles, I want access to data that has helped generate those articles, and the citation data of those articles as well as the journals they are published in. I want access to images or illustrations and their metadata that might be found in those articles. I want access to the data about the authors, collectors, expeditions, institutions that help facilitate the production of those articles. To me these are the "research outputs" I'm interested in.

[Slide 3]
And my definition of what constitutes "access" is just as wide. Access to me is synonymous with reuse. I don't just want the ability to look at or read the research content. Open Access is having the ability to reuse that "research output".

In this presentation I want to illustrate how important this generous view of "Open Access" is to what I do. I'm going to take you through three different workflows. I'm aiming to show you how I can only do what I do if institutions and individuals provide Open Access to biodiversity knowledge.

I want to convince you to think about Open Access as I think about it. With a wide definition of "research outputs" and with "access" meaning the ability to reuse those research outputs.

[Slide 4]
Now my playground is the Biodiversity Knowledge Graph - the interconnected network of taxa, taxonomic names, publications, people, species, sequences,

images, and collections. I work to enrich this network by adding to it and making those connections. And the tools I use to play in this particular sandpit are Wikipedia, WikiCommons and Wikidata. I use these three Wikiprojects to curate and link information and data on biodiversity with and to each other. Hopefully helping to enrich the knowledge graph and in turn making it easier for others to build upon it.

[Slide 5]
Now I'm sure all of you are aware of Wikipedia - the online encyclopedia editable by anyone. But perhaps you aren't aware of the other two Wikiprojects I've mentioned. WikiCommons is an online repository of free to use images, sounds and other media. It is full of public domain or openly licensed images that can be reused for free by anyone. I upload images into WikiCommons to reuse on Wikipedia, Wikidata AND on other websites.

Wikidata is a multilingual linked open database. It is a structured database that can be edited, queried and reused by anyone. Wikidata is CC0 licensed.

Importantly, Wikidata isn't just an open database, it also an identifier hub - that is it links out to other databases. I add data and identifiers to Wikidata. These data and identifiers can then be queried and reused by anyone for any purpose.

So how do I use these three Wikiprojects in conjunction with other open access content for my work?

[Slide 6]
Well my first example is my ongoing New Zealand endemic moths project. I got inspired to work on New Zealand endemic moths because of Manaaki Whenua Landcare Research, a New Zealand Crown Research Institute.

[Slide 7]
They had started a citizen science project, getting schools to participate in studying and generating data on New Zealand moths. It seemed obvious to me that Wikipedia would be one of the first places the children would look for information on the species they were studying.

[Slide 8]
But many of the approximately 1650 endemic moths of New Zealand lacked a Wikipedia page. So I started creating or expanding articles. The first place I look to when researching articles is the publication that contains the original scientific description of the species. It is MUCH easier to study a species if you know the

name of it and can distinguish it. The original scientific description is the start of this process. So I wanted to make sure that the original description was in Wikipedia.

And the place I first go to to try and get these original scientific descriptions is the Biodiversity Heritage Library. BHL is a consortium of natural history and botanical libraries that cooperate to digitize and make accessible and reusable the legacy literature of biodiversity**.**

And because BHL contains open access content I can not only cite those scientific articles, I can copy the original description of moths into Wikipedia, making them easier for citizen scientists to find.

[Slide 9]
While writing the wikipedia article I'll also attempt to find an image of the species. Obviously, if you've got citizen scientists looking for endemic moths it really helps if they have some sort of idea what the moths might look like. For NZ endemic moths, I've got several open access datasets of images to choose from. Both Manaaki Whenua Landcare Research and Auckland Museum license their images under a CC BY license - an open license that only requires attribution when you reuse those images. So I was free to download those images and upload them into WikiCommons for reuse, both in the Wikipedia article and in the related Wikidata item for that species.

I think this is a good example of how an institution's open access policy for their research content can attract folk to actively engage with and reuse their content.

[Slide 10]
But with New Zealand having so many species of endemic moths, many do not have an openly licensed photo available. If I can't find a reusable photo, I go back to the scientific literature to see if there might be an illustration. If the literature is truly open access I can download that illustration and add it to Wikicommons for reuse.

But often for a lot of NZ endemic species, it is normal to find that there are no openly licensed photos or illustrations available. This can really hold back citizen scientists in gathering biodiversity data. How can data be collated if folk don't know what they are looking for?

I've gotten so frustrated at this lack that I've started directly contacting institutions with collections, campaigning and persuading them to release their species images more openly. One of my success stories is New Zealand's national museum Te Papa Tongarewa. They are currently in the process of clearing their specimen images to be reused under a CC BY copyright license. This will give me one more open access institution to check to see if they have appropriate specimen images.

[Slide 11]
While still writing my wikipedia article, I'll also double check that the "taxonbar" is included in the article. This taxonbar is where identifiers sourced from external databases are displayed in the Wikipedia article. The taxonbar automatically pulls these external database identifiers from the Wikidata species item. Obviously it relies on the identifiers from those databases being added to Wikidata. And of course adding this data to Wikidata is made so much easier if the sourcing institution's dataset is Open Access - openly licensed for reuse. It can then be easily ingested into Wikidata and used to populate appropriate wikipedia articles.

[Slide 12]
Now it is possible that an institution may have a whole database that isn't as generously or as openly licensed. But individual facts and identifiers are not in themselves copyrightable. So while Wikidata editors may not be entitled to copy the entire dataset in bulk, editors are able to add identifiers individually.

[Slide 13]
For my purposes the two particular databases I want to ensure are linked to a species Wikidata item are the iNaturalist and Global Biodiversity Information Facility (GBIF) identifiers. iNaturalist is a citizen science website and app that enables anyone to collect species observations. iNaturalist also happens to be one of the best places to get images of NZ endemic species.

BUT it has an issue. It's default license is the Creative Commons Attribution Noncommercial use. The data and images contained in iNaturalist aren't automatically Open Access. Unless the user deliberately changed their default license I'm unable to reuse those species images in WikiCommons.

I've been working with other Wikipedia editors to raise this issue with iNaturalist and to attempt to persuade them to change their default license. This is still a work in progress. In the meantime I've been reaching out to individual iNaturalist users, requesting they change their default license to a more open license, allowing the open reuse of their images. This can pay real dividends even if I just convince a few, as sometimes their images are the ONLY images of these species available.

The other identifier I'm keen to link to a species Wikidata item and in turn to the species Wikipedia article is the GBIF identifier. The Global Biodiversity Information Facility (GBIF) is an international network and research infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth. It’s doing it’s best to provide open access to data. I say “doing it’s best” as some institutions that supply datasets use closed reuse licenses.

The reason I want to link these two identifiers to Wikidata species items is to ensure that these data are interconnected. I’m a firm believer that the more connected data on a species is, the easier it is for people to find, reuse and query that data. And in turn the easier it will be to build upon that data. Once both the iNaturalist and the GBIF identifiers are added to the species Wikidata item, those identifiers will also be pulled into the taxonbar on the Wikipedia article. Readers of that article can then just click on the identifier link to be taken directly to either of those databases to find out more on the species.

[Slide 14]
While in Wikidata I’ll also ensure other information and data is added to the species item. Information like the endemic nature of the species. The author and year of publication of the original scientific description. I’ll link to the article that contains the original description.

This enables anyone who is interested, such as museum curators or taxonomists, to easily find these scientific publications at just a click of the mouse. And I’ll cite scholarly articles or published datasets to support the factual statements I’m making in Wikidata.

[Slide15]
So to summarise, I’ve created an article that includes a reference to the original description. I’ve then added at least one image to Wikicommons, and reused the image in the Wikipedia article and the species Wikidata item. I’ve created or expanded the species item in Wikidata, added as many identifiers from different databases as I can find, along with other information such as the author and year of description, endemic nature, hosts of the species. I’ve added references supporting these statements.

All this work to help citizen scientists and others to easily get the information they need to learn about and identify the moth species they are trapping and gathering data on. So that more knowledge on these species can be generated.

And ALL this relies on Open Access to information, data and images. Hopefully you are beginning to realise how much work can be done linking information in the Biodiversity Knowledge Graph if the content and databases created and provided by institutions and individuals are open access. But I’m not finished yet.

[Slide 16]
I return to the iNaturalist website. Because Wikipedia is itself open access, that is licensed for reuse under a Creative Commons Attribution Share Alike license, iNaturalist can automatically ingest the Wikipedia article I've written into its site.

[Slide 17]
iNaturalist allows its members to upload the species images previously added to WikiCommons into iNaturalist. So the Open Access images I've sourced from institutions such as Manaaki Whenua Landcare Research and Auckland Museum can be added to iNaturalist species pages. This is important as these images of species have been accurately identified by knowledgeable experts and can be therefore used by citizen scientists to compare against their own observations.

If I've done my job a citizen scientist using iNaturalist will have both expertly identified images and the Wikipedia article to help them identify their own observations of that species. So Open Access content will also help ensure the accuracy of citizen science generated biodiversity data.

[Slide 18]
The citizen scientists upload their own species observations into iNaturalist. Once that identification is confirmed to research grade, that observation data is ingested into GBIF. Remember the GBIF external identifier is listed in Wikidata, and is in turn is included in the taxonbar in the Wikipedia article.

So any confirmed observation data on a species in iNaturalist adds to the quality of the data linked to in Wikidata and Wikipedia. If I've convinced the iNaturalist user to change their default license, to allow open access to their image, the image of their observation is also available for upload in Wikicommons.

In this way a virtuous cycle of reuse has been created, each addition building on the last. NONE of this would have been possible for me to do without Open Access principles being put into practice by a wide variety of institutions and individuals.

[Slide 19]
Now the second example of how Open Access empowers my work is a more digital humanities example but with a Biodiversity Knowledge twist. Because as well as being interested in New Zealand endemic moths I'm also really interested in Women Scientific Illustrators.

My aim with this example is to show you how it is possible to reuse open access data and images to help encourage further research into women scientific illustrators. This helps give attribution and credit to these underrepresented and often unacknowledged contributors to scientific knowledge.

[Slide 20]
Now I got interested in women scientific illustrators BECAUSE of open access content. I was volunteering for the Biodiversity Heritage Library, the open access digital repository of biodiversity literature.

[Slide 21]
BHL has an extensive collection of scientific illustrations in Flickr, sourced from their literature. I would volunteer by tagging those images with taxonomic names as well as with illustrator tags. This tagging made those images easier to find in Flickr for reuse. But most importantly to me this tagging was also incorporated into Wikicommons when those Flickr images are bulk uploaded by other wikicommons editors.

I was collaborating with another volunteer Michelle Marshall on this tagging work. While doing this work both Michelle and I were enthusiastically encouraged by Grace Costantino, the BHL Outreach and communication manager.

[Slide 22] - So many women
While tagging we would come across women artists. So many women. Amazing women, about whom there appeared to be little known or written. Some of these women would illustrate multiple articles, books and scientific publications. Others were writing books and articles, massing collections of specimens, having species named after them as well as publishing those scientific illustrations.

Both Michelle and I were keen to find out more about these women but there was often very little written about them on the internet. Every once in a while there would be a woman who had significant coverage, enough so that a wikipedia article had been created for her. But this was the exception to the rule. This lack of coverage was frustrating to both of us.

[Slide 23]
So both Michelle and I started researching. Me with the aim of writing wikipedia articles, her with the aim of writing blog posts and enriching her own scientific art social media accounts.

[Slide 24]
The women who created scientific illustrations didn’t tend to exhibit in art galleries. Their art was created to enhance scientific publications and wasn’t treated as stand alone work worthy of critique and public display. It was often a real challenge to find information about these women.

Historically much of these women’s illustration work was not regarded at the time of their creation as being worthy of comment. At most they received a passing remark in reviews of the publication or acknowledgement by the author of the work. This lack has resulted in them being overlooked by library catalogers. Often they and their contributions were simply not recorded in databases.

[Slide 25]
But working together, Michelle and I undertook to help rectify this.

Often we could track down enough information to work out who these women were, what scientific works they had contributed to and whom they had worked for. One of the most useful resources for this is of course the Open Access diamond that is BHL. While we were doing this work, BHL enabled a full text search of their content. This significantly improved our ability to find these women within the scientific literature. Sometimes having Open Access to literature is of little use if the technology doesn’t support the finding of the content inside it. That ability to full text search BHL content revolutionised our research work.

[Slide 26]
I would then add the women, their data and any identifiers to Wikidata. Wikidata is a resource, just like Wikipedia, that informs Google’s knowledge graph. So if I can get these women into Wikidata, I would go some way to ensuring that they and their data is findable by other researchers via a Google search.

I would also use the reference section of the wikidata statements not just to provide evidence in support of the statements themselves, but also with an eye to helping collate all the links we’d discovered during our research. I want to leave a research trail making it easier for me or others like me to find these sources and to then reuse them for example in wikipedia articles on these women.

Obviously if external identifiers did exist I also wanted to include them. To my disappointment, despite the prestige of the works these women were illustrating, many of these women were not listed in external databases. I always check VIAF, the Virtual International Authority File database that gives information sourced from national libraries.

Although VIAF would frequently list the author of the scientific publications, information on the illustrator of these works were often missing. Libraries prioritised those who wrote the words rather than those who created the art that illustrated the work, even if the illustrations made up a large proportion of the publication.

[Slide 27]
I would also check the Stuttgart Scientific Illustrators database. This is one of the most comprehensive databases for scientific artists. Sometimes the women would be in this database but sometimes not. Sometimes only under their maiden name or only under their married name, not both. Or if both they might be listed as two seperate people. Although a fabulous starting point, this database wasn't as comprehensive as I needed.

But the wonderful thing about this particular database was how responsive it's creator, the History Department of the University of Stuttgart, is to emails. Both Michelle and I write to them including our research on particular women illustrators asking for these women to be included.

They would add these women to their database and generate an external identifier. They are also able to link in resources that neither Michelle nor I had access to. Often more data was added on these women in the Stuttgart database as a result of their further research.

Now the Stuttgart Scientific Illustrator Database as a whole is not Open Access in that the whole database is copyrighted. But as explained previously individual facts and identifiers, are not copyrightable. So I can add their specific identifier to the artist's Wikidata item. I can also add statements about the facts contained in the database, using the database as a reference to support the statements I'm making in wikidata.

[Slide 28]
Michelle and I would also contact BHL about these women. We would request that the BHL catalogue record be amended to include them. If the necessary criteria was satisfied, BHL would edit the metadata and in doing so creating another external id - BHL creator id. This identifier would help collate and connect the works the women illustrated to the women themselves. And the identifier can be added to Wikidata.

Obtaining a creator id can ensure a cascade of linked open data. This can raise the visibility of these underappreciated women to researchers and make it easier for researchers to find the works the women contributed to.

Slowly I began to feel we were making a real difference in surfacing these women. At least now when folk googled them, the wikidata item would likely make an appearance in the search results. The images the women had created, that had been uploaded into Wikicommons, would be shown in the Google image search results. Our research, tags, blogs, wikidata items and the external identifiers created as a result of our requests were all coming together making these women easier to discover.

[Slide 29]
But our work really came to the fore when the BHL held their “Her Natural History” campaign. All posts in Her Natural History

This was a multi institutional, multi platform campaign to raise awareness and to celebrate the contributions of women to natural history. This campaign resulted in numerous outcomes many of which had a direct impact on the richness of metadata available on these women.

[Slide 30]
The BHL Cataloging Group added more female contributors to the BHL catalogue generating more external identifiers.

More artworks by these women were added to the BHL flickr feed. These were all Open Access, that is either in the public domain or were openly licensed for reuse and therefore were able to be uploaded into Wiki Commons.

Numerous blog posts were written by employees of the BHL member institutions. Some of these blogs used the research Michelle and I had undertaken as a starting point, picking it up and ran with it. Their research often resulted in the discovery of new sources of information that assisted in ensuring these women obtained a wikipedia article.

During the campaign there were also three Wiki Workshops. These events added significantly to the work I was undertaking to make these women, their illustrations and sources describing them more easily findable.

I really believe that this BHL campaign shows how Open Access to resources can inspire work that in turn generates more knowledge of and about a subject. Open Access becomes the leaping off point enabling research to be collated and linked and sparking further research assisting these overlooked women to get the recognition they deserve for their work.

[Slide 31]
Finally my third example. With this I’m hoping to show how Open Access to data and content can help bring attention and professional recognition to specimen collectors, particularly historic women specimen collectors.

Currently Natural history collections, their curators and collectors have difficulty in finding ways to illustrate the importance and impact of their work. Unlike scientific publishing where citation metrics help illustrate and reward expertise, there are few metrics available that show the skills needed to collect and identify specimens, maintain them, digitize their labels, create and enhance natural history data.

[Slide 32]
BUT there are websites and tools that are being developed that use open access data sourced from Natural History collections to overcome this issue.

One site that does this is Bionomia Tracker. This is a website that has been developed to help provide metrics and visualisations to ensure the numerous people making significant contributions to science through their collecting, curating or identifying of species get the same recognition as those scientists who do the writing of scientific papers describing those species. Bionomia Tracker was specifically created as a way to help acknowledge the importance of this type of work.

Bionomia tracker uses either the contributor’s ORCID ID or, if they are deceased, their Wikidata item, to link the collector to their collections. It uses data that natural history organisations have shared with the Global Biodiversity Information Facility (GBIF) to help make this connection.

Natural history collections ADD specimen datasets to GBIF. These datasets list the collectors of the specimens. Bionomia Tracker links those specimens to the collectors ORCID or Wikidata item.

GBIF also tracks the scientific papers that reuse and cite those specimen datasets. Bionomia Tracker uses that citation data to link those papers to the specimens and in turn to the collectors of those specimens. This linking surfaces the scientific impact of those specimens and helps to illustrate how important this collection work is.

[Slide 33]
In this way Bionomia Tracker gives metrics that show the impact of a person’s collecting work.

[Slide 34]
Now when I volunteer for Bionomia Tracker helping to make these connections I tend to concentrate on historic women collectors. There are numerous under acknowledged women who have contributed to scientific knowledge by collecting specimens. These collections continue to be held and studied in museums or herbariums. As more and more of these collections are digitised and their datasets are published in GBIF, more and more historically significant women collectors can be recognised via Bionomia Tracker.

This information is also a rich vein of data on early women scientists particularly as at that time when they may have been reluctant or unable to obtain academic positions, publish scientific works or join scientific societies or clubs due to the social norms of the day.

[Slide 35]
But in order to link these women to their collecting via Bionomia Tracker the first thing I have to do is get these women in Wikidata. I've often only got a name on a specimen label to work from. I undertake research. I'm attempting to find the woman's birth and married names, their date of birth and death, where they lived, where they collected and where their archives and collections have been donated. I'm also aiming to find a reusable image of them as well as a sample of their handwriting to assist others to identify their specimens.

I want to add all this information to Wikidata as I have to disambiguate the person I'm researching. I want to be sure the specimens are being attributed to the correct collector.

[Slide 36]
Much of this research work relies on my ability to have open access to a wide variety of heritage content. Open access to information and documents such as birth, death and marriage certificates, passport applications, genealogy websites and databases, town property records, high school and university yearbooks, archive records, field books ... the list goes on.

So I tend to rely heavily not just on Open Access content sourced from the BHL but also sites such as the Internet Archive that can link to year books and archive records, as well as genealogy websites that link to open access primary sources such as birth certificates or passport applications.

[Slide 37]
Once I’ve added all the necessary data and identifiers to the woman collector’s wikidata item, Bionomia Tracker can automatically ingest the information in wikidata. Bionomia Tracker can do this because Wikidata is itself open access. I also have the option of manually adding the Wikidata item, if it fails to meet the criteria for automatic ingestion.

Once the woman is in Bionomia tracker volunteers can help those women “claim” their collections, enriching not just the linked open data about the specimens themselves but also ensuring these women get credit for their vital work.

And then Bionomia tracker identifier can also be added back into Wikidata making another virtuous cycle of reuse.

[Slide 38]
Adding these women to Wikidata also has other benefits. The Wikidata item can link them to scientific literature they have written, to information held on them by archives, libraries and museums or to the species that have been named after them. And all this information can be queried leading to new discoveries about these women.

But the finding, linking and reusing this information relies on Open Access to data about these women. With more and more Libraries, Archives, Museums, genealogy databases, Government records and documents are becoming open for reuse, the easier it is for me to find and reuse information and data on these women. Open access empowers me and others like me to link them to their work and help obtain recognition for their valuable contributions.

[Slide 39]
I’m hoping that by sharing the above three workflows I’ve shown you how important it is to take a wide view of what Open Access means. Open Access to published scientific articles is only the tip of the knowledge iceberg. I want you to view Open Access the way I do - as the ability to reuse knowledge. Without the ability to reuse images, data, citation data, authority control links and so much more, I and others like me wouldn’t be able to do what we do.

And finally I want to challenge you to bring that Open Access mindset to everything YOU do - your presentations, your photos, your articles, even your tweets. Open access is MY default, I hope you’ll consider making it yours too.

[Slide 40]