BioHackrXiv

Network analysis of specimen co-collection

Sofie Meeus[1], **Tom August**[2], **Lien Reyserhove**[3], **Maarten Trekels**[1], **and Quentin Groom**[1]

1 Meise Botanic Garden, Nieuwelaan 38, 1860 Meise, Belgium **2** UK Centre of Ecology and Hydrology, Maclean Building, Crowmarsh Gifford, Wallingford OX10 8BB, UK **3** Instituut voor Natuur- en Bosonderzoek, Team Oscibio, Havenlaan 88, 1000 Brussel

BioHackathon series:
BioHackathon Europe 2021
Barcelona, Spain, 2021
BioHackrXiv

Submitted: 26 Nov 2021

Published by BioHackrXiv.org

Introduction

Biodiversity data are collected by people, and those people often work in teams. Those teams may be large, particular when they are part of an expedition, though they may be also be as small as two people. Data on who works together can help refine biodiversity data in many ways. It can help us cross-reference data to ensure it is consistent and valid (Groom et al., 2020). It helps us acknowledge the contribution to science of all of the participants. It helps us understand how scientific collection, learning and communication operates and this can give us insights into the biases and effectiveness of the collection process. Also, the relationships between people, and the organisations they are members of, are interesting from a historic and sociological perspective.

Network analysis has been used for some time to explore the relationships between people, but the connections analysed may be strong, as in the case of citation networks where the cooperation between people is long term, or weak in the case of Twitter analytics, where only a mouse click connects people. Here we specifically analyse the co-collection of biological specimens by people. Co-collecting of a specimen requires that those people involved in the collection process travel, organise and explore together. Therefore, one could argue that such a connection could be even stronger than co-authoring a publication, though doubtlessly the degree of engagement varies considerable. Networks based upon co-collection have been created before for specific groups, for example to analyse the botanical exchange clubs of the United Kingdom (Groom et al., 2014) and for a specific herbarium (Siracusa et al., 2020). However, in this paper we approach the co-collection networks from a global perspective using the data from Bionomia. Bionomia is a community based project that allows users to associate stable identifiers for people, such as ORCID and Wikidata Q numbers, to the anonymous text strings transcribed from specimens in museums and herbaria.

The aim of this project was twofold, firstly we wanted to profile the co-collectors, how co-collecting has changed over time and who people co-collect with, and secondly we want to examine whether co-collecting can be used to reveal errors in the data and is therefore a means of validation. Still, as we started examining these data we realized their value in discovering the contribution of women to collections. Women's contribution to natural history and taxonomy has been underacknowledged and we have therefore examined that aspect in more depth (Lindon et al., 2015).

Methods

Data sources

Data on collectors were downloaded from the Bionomia website (2021-11-06). This comma separated file contains three columns (Subject,Predicate,Object), the URI of the GBIF id of the specimen, the identifier of the Darwin Core term (recordedBy or identifiedBy) and the person identifier (ORCID or Wikidata Q number). This file was imported into a table in an

SQLite database (Hipp, 2020). All rows referring to identifications of specimens were deleted, leaving only those related to specimen collection. A query was then run using a self-join on specimen ID to create a new table containing two rows with pairs of collectors that collected with each other. This table was then exported and the number of specimens was calculated per collector pair, to create a file of network edges with the pairs of collectors and a weight based on the number of specimens they had in common.

Demographic and gender information on the collectors was retrieved from Wikidata using the notebook 'get_collector_gender.ipynb' (see Code and data section below). Using the SPARQL endpoint of Wikidata, the script collected the relevant information if it is available. ORCiD records do not contain gender or demographic information. Therefore we can only retrieve the information if those people are represented in Wikidata and their ORCiD is present in their Wikidata entry. The output of the script creates a nodes file containing the ID of the person and columns for the gender and demographic information.

Using the network edges and the nodes list, the Jupyter notebook 'age_differences.ipynb' starts with filtering out the unique interactions between people. Using the demographic information, it was possible to derive the age differences between the interacting people. This can serve as a metric to filter out errors. It is possible to detect wrong assignments of people to specimens, because the age gap between people is either impossible or at least highly unlikely. Since in many cases gender could also be retrieved, it also enabled the analysis of gender as parameter in the network of collectors.

Collector network visualization

To visualize the network the nodes and edges files was imported into Gephi (Bastian et al., 2009) as an undirected network with weighted edges. The network was laid out using the Yifan Hu algorithm (Hu, 2011). The weight of the edges was equal to the number of specimens a collector pair collected together (not shown in Fig. 1).

Results

The network of collectors

The network contains 3009 nodes and 4330 edges (Fig. 1). The average degree is 2.88, meaning that the average collector has about three co-collectors. Note that for this analysis we included only collectors that had at least one co-collector, collectors with a degree of zero were excluded from the analysis. The average modularity of the network of 0.84 is quite high as can be seen in Fig. 1. Modularity in this particular method for community detection (Blondel et al., 2008) ranges between -0.5 (non-modular clustering) and 1 (fully modular clustering) in which all the edges fall within the communities. This analysis identified 327 communities. The top five of largest clusters with their characteristics is shown in Table 1. The total diameter of the network is 22 and the average path length 6.99 (Brandes, 2001). It is notable that these networks include some well known historical collectors, including Carl Linnaeus and Alfred Russel Wallace. The network correctly identifies Olof Celsius and Henry Walter Bates respectively as their co-collectors (Fig. 1). The data gathered from Wikidata also includes the gender of the person, therefore we are able to analyse the network for gender differences in co-collecting. This has been visulized in figure 2.

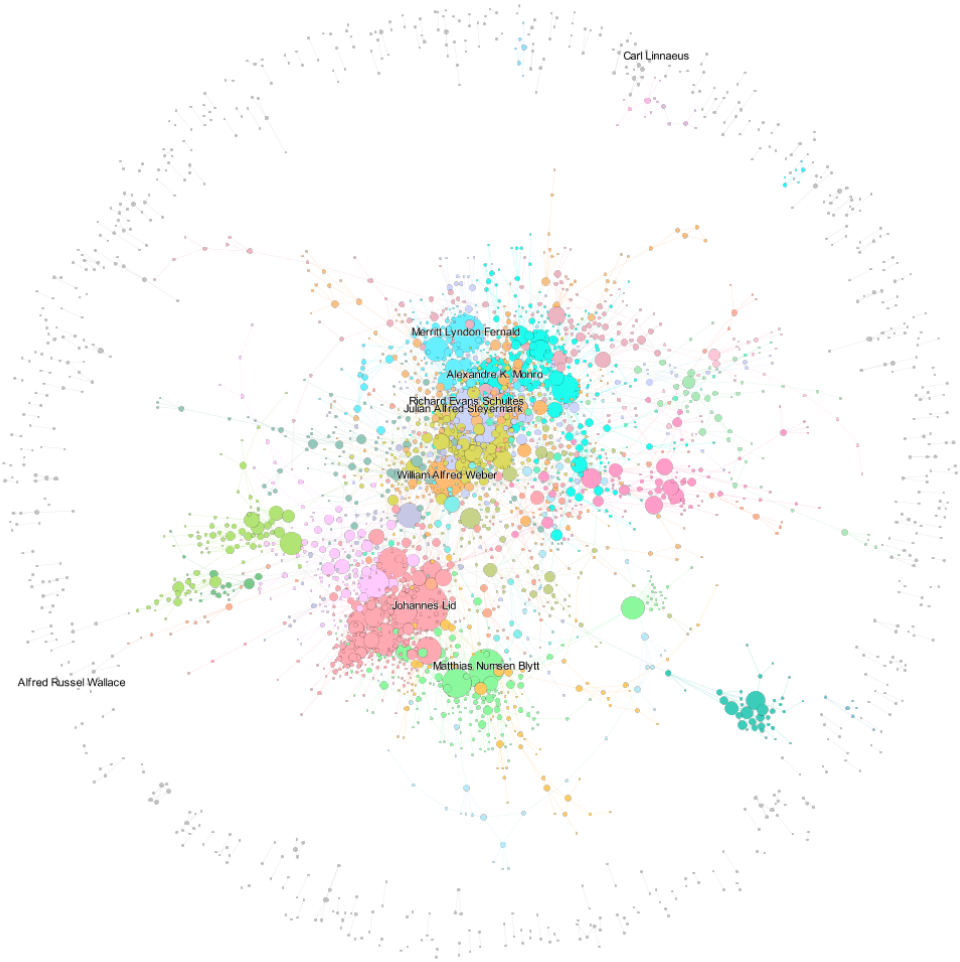


Figure 1: The network of collector collaborations for specimens identified in Bionomia (https://bionomia.net/). This was created in Gephi (Bastian et al., 2009). The size of the nodes is determined by the degree of the node (i.e. number of people they collected with) and the colours of the nodes is determined by a community detection algorithm and coloured for the largest modules within the network (Blondel et al., 2008)

Table 1: Summary of the collector network in terms of the top five largest clusters with their number, colour, size (number of collectors) and their node with the highest degree and the name of the person represented by this node.

number	colour	nodes	highest degree	person with highest degree in cluster
90	Salmon	259	45	Johannes Lid
121	Turquoise	184	34	Alexandre K. Monro
11	Yellow-gold	163	27	Julian Alfred Steyermark
65	Violet	155	24	Richard Evans Schultes
45	Green	153	34	Matthias Numsen Blytt

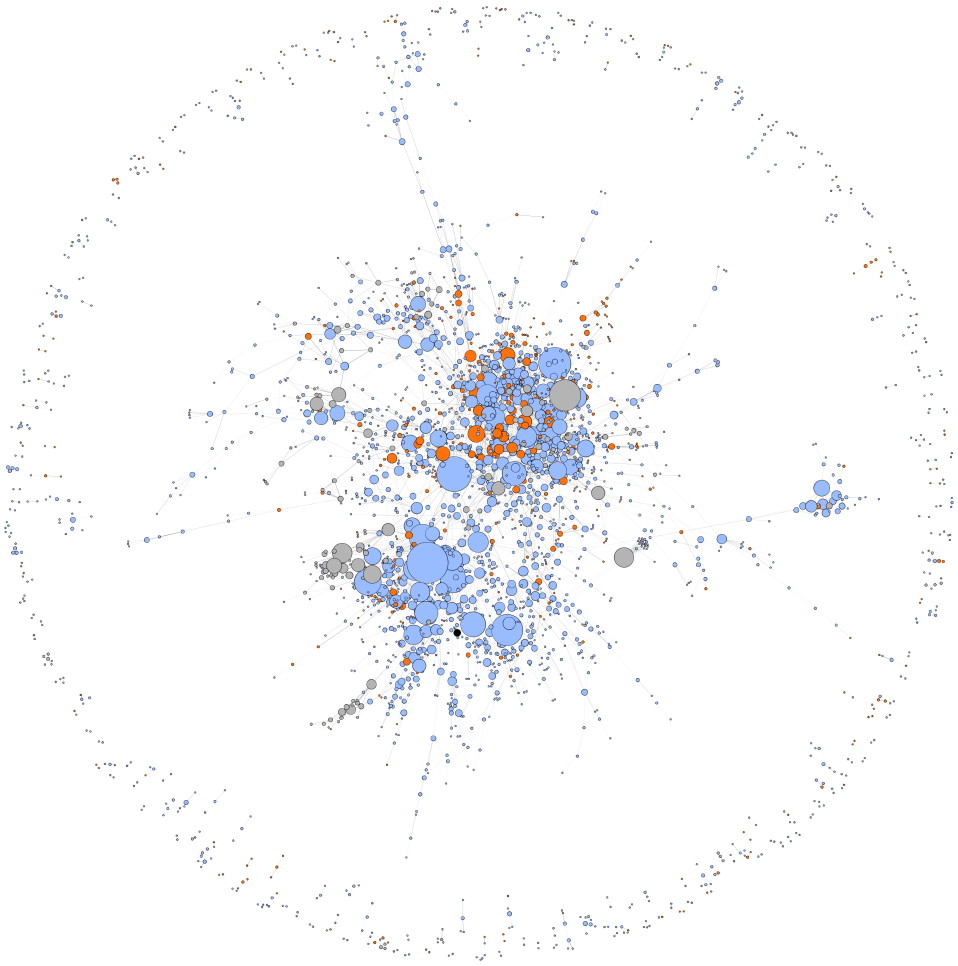


Figure 2: Gender of collectors visualized in the collector network with node size proportional to the weighted degree (i.e. number of people they collected with). Colours: orange=women, blue=men, grey=unknown. The black node is that of Hanna Resvoll-Holmsen (see below)

Most people collaborate with only with one person, but there are some super-co-collectors who collaborate with many people. The top three men and women with the highest number of collectors is listed in table 2. However, men tend to have many more co-collectors than women (Fig. 3).

Table 2: The top three men and women with the largest number of co-collectors, ordered alphabetically by their surnames

Wikipedia	Wikidata
Elizabeth Gertrude Britton	Q2567402
Merritt Lyndon Fernald	Q2656885
Johannes Lid	Q94522
Alicia Lourteig	Q454806
Elisa G. Nicora	Q5829538
William Alfred Weber	Q4105706

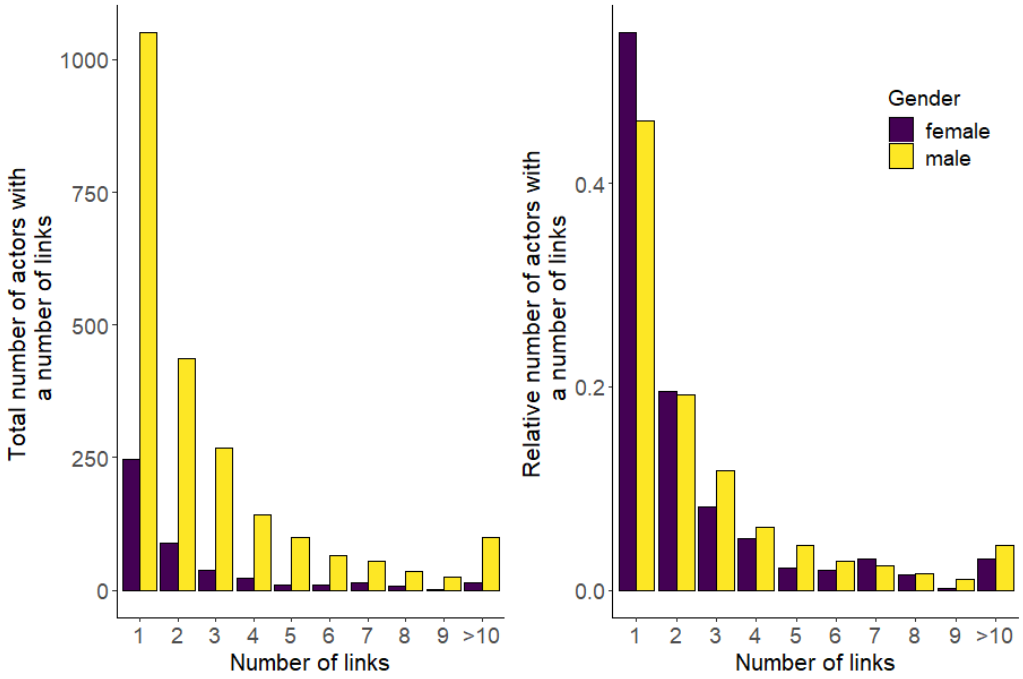


Figure 3: The absolute (left) and relative (right) number of collaborations for women and men who collected specimens identified in Bionomia (https://bionomia.net/)

After analysis through the age differences notebook, a histogram of the age differences could be constructed for the different gender combinations (Fig. 4). The distribution of age differences suggests that the cut-off of realistic differences in ages is around 50 years. Also intuitively it seems reasonable to assume that it is worthwhile checking the records that show a bigger difference in age. The fraction of edges that should be investigated further is around 5.5%. This is a significant number of records that could be fedback to Bionomia to be checked and corrected by the community.

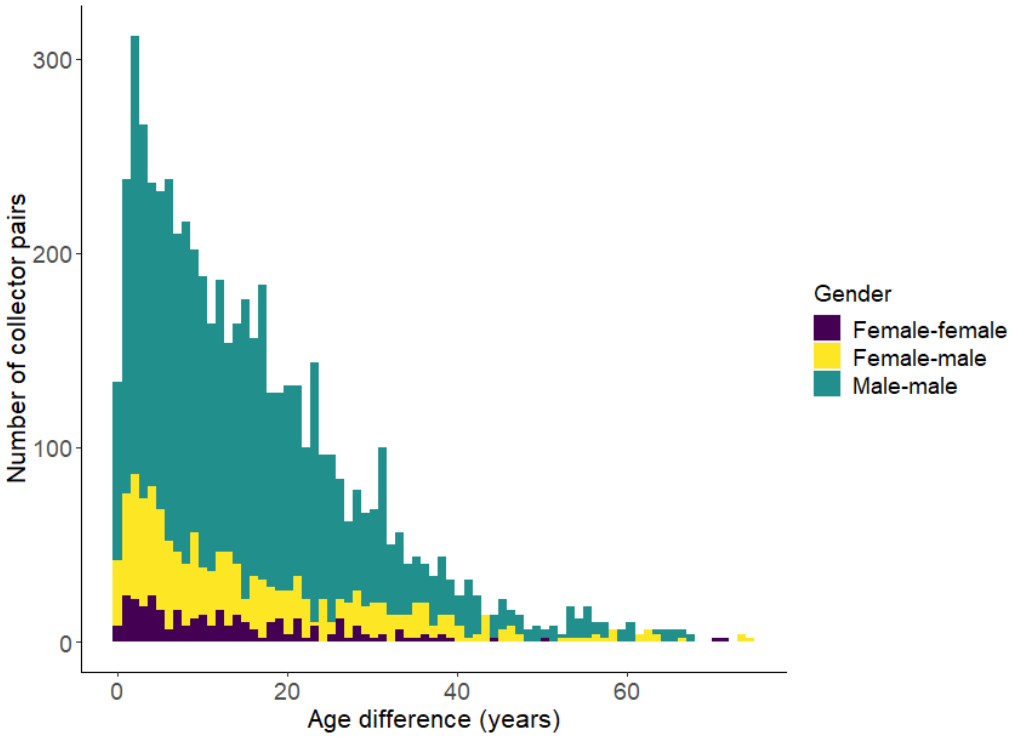


Figure 4: Histogram of the age difference between people. The results are shown for each of the different gender combinations that could be derived from the data

The number of co-collected specimens collected increases with time, as might be expected, specimens in general have increased with time. However, the earliest co-collections are purely male-male combinations. In the late 18th century the first specimens were collected by mixed gender pairs and it takes until the beginning of the 19th century that female only collecting pairs appear in the data. In the 20th century, the pure male collecting teams are dropped below 70% and are still decreasing.

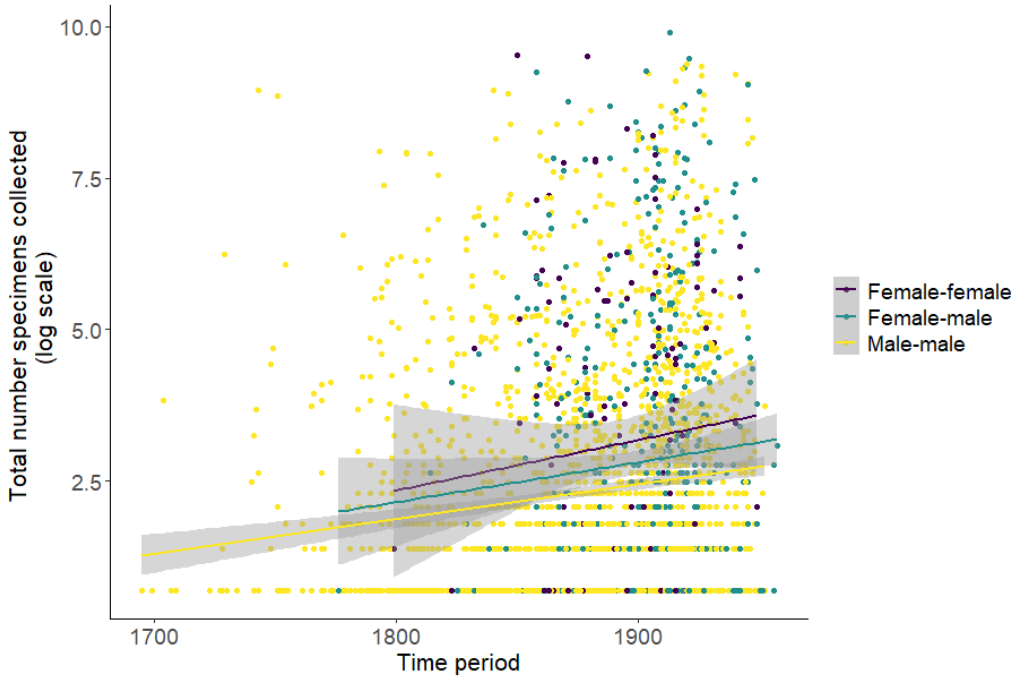


Figure 5: Number of specimens collected as a function of birth date of one of the people. The colors are indicating the combination of genders

Discussion

Given that there are around 2 billion specimens in natural history collections worldwide (Ariño, 2010), and only a small proportion of those have been fully digitized and linked to identifiers for their collectors, these networks provide an incomplete view of the whole co-collection network. In reality the network has many more people involved and many more co-collected specimens. Nevertheless, it is clear that patterns are emerging of a highly connected network that spans time and geography. Though women are a relatively small component of these networks, their importance is increasing and there are some exceptional women who had a major contribution.

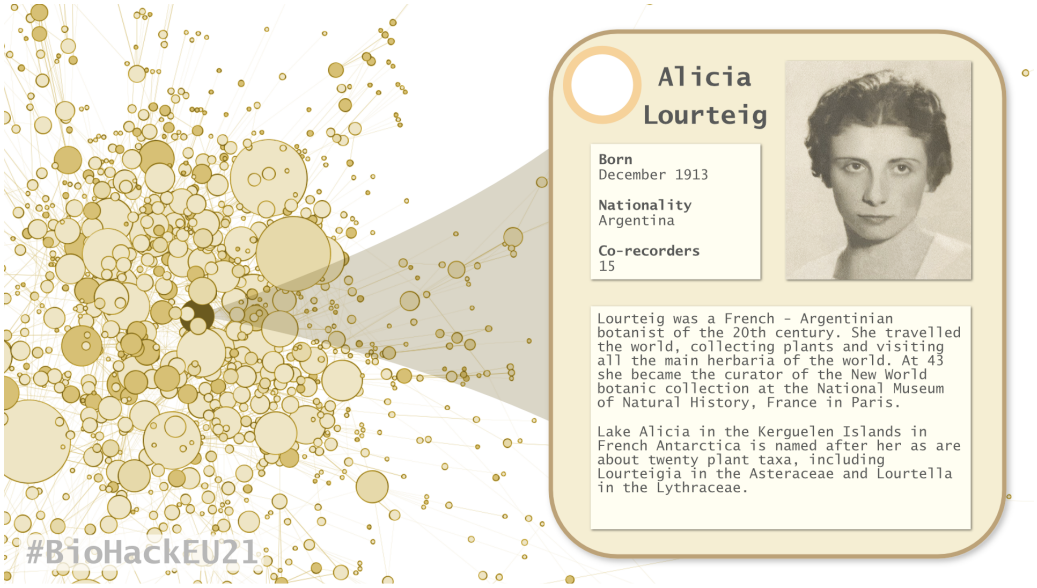
It is worth noting that we only looked at co-collecting pairs, while specimens can be collected by larger teams than two. A more sophisticated analysis should consider this in more depth. One approach is to weigh the co-collection less if it was collected by more people. This is to take into consideration that the strength of the interaction is likely to be weaker in larger teams (Siracusa et al., 2020). It should also be considered that in Bionomia people are associated with specimens as individuals and not simultaneously if there is a team. Therefore, many co-collector combinations are yet to be found and some collector teams will be incomplete. This is unlikely to have an impact on our conclusions, however, before a more in-depth study is completed a gap analysis is needed to reveal the extent of these gaps. A perhaps larger bias in our results is the preferences of the Bionomia users and the availablity of digitized collections on the Global Biodiversity Information Facility. Both of these factors will lead to geographic and perhaps gender biases in the data and only with further digitization and disambiguation will this be resolved.

We started this project aiming to identify errors in specimen data through co-collection. We have partially achieved this by profiling the age difference between collectors. However, there is much more that could be done. We did not explore linking these networks to the locations that specimens were collected. People cannot be in the same place at the same time and can therefore not co-collect if they were in different places. Similarly, by connecting these networks to the specimen data we can look for other kinds of outliers, such as mismatches in the

taxonomic interests of the co-collectors. Furthermore, by making links between co-collection, geography and taxonomy we can perhaps reveal more about the lives and interests of poorly documented collectors through their association with the better documented ones. This might be particularly true for the women in the network.

Outreach

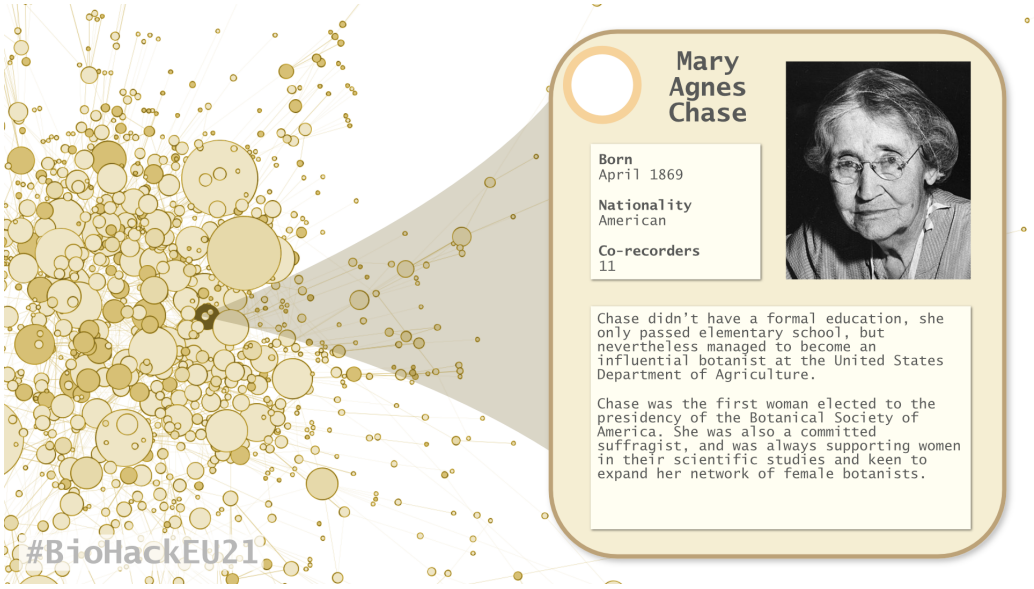
Given our results on women in science we thought we would take the opportunity to highlight the work of three of the highly connected female collectors. Below are the infographics we made and the texts we tweeted on these women. More information can be found about these women in these publications (Fosberg & Swallen, 1959; Fuglei & Goldman, 2006; Sastre, 2003).



Tweet: French–Argentinian botanist Alicia Lourteig has about 20 plant taxa named after her.

Tweet: Hanna Resvoll-Holmsen was a pioneer conservationist and ecologist, advocating for the preservation of natural ecosystems in her paper Om betydningen av det uensartede i våre skoger

Her publication we mention is in the references list (Resvoll-Holmsen, 1932).



Tweet: Mary Chase was the first woman elected to the presidency of the Botanical Society of America. She was also a committed suffragist, and was always supporting women in their scientific studies.

Also the tweets included the hashtags #WomenInScience #STEMWomen & #BioHackEU21

Future work

As more data becomes available there is considerable scope for repeating and expanding this project. We do not anticipate the general trends to change, but we will be able to study the network, and its various sub-networks in much more detail. It may also be valuable to compare these collection networks with citation and co-authorship networks. Networks, such as these, will help us understand the provenance of collections, and the biases they contain, thus improving the overall metadata of collections. It would also be useful to collaborate with historians, museologist and social scientists to get a different perspective on what these networks tell us about people and the collection process (Fyfe, 2006; Lourenço & Dias, 2017; Mignan, 2018).

Code and data

All code and data can be found on GitHub.

Author contribution statement

QG had the concept, acquired funding and supervised the project; SM, LR & MT curated and validated the data and conducted the formal analysis, LR & MT wrote the software; SM and TA created the visualizations; SM determined the methodology and conducted the investigation. All authors contributed to the writing, reviewing and editing of the paper.

Acknowledgements

The authors thank the organizers BioHackathon-Europe for their support and excellent organisation. This work was also facilitated by the Research Foundation – Flanders research

infrastructure under grant number FWO I001721N and the BiCIKL project of the European Union's Horizon 2020 Research and Innovation action under grant agreement No 101007492.

References

Ariño, A. H. (2010). Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, *7*(2). https://doi.org/10.17161/bi.v7i2.3991

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks*. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, *25*(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249

Fosberg, F. R., & Swallen, J. R. (1959). Agnes Chase. *Taxon*, *8*(5), 145–151. https://doi.org/10.2307/1216753

Fuglei, E., & Goldman, H. V. (2006). Hanna Marie Resvoll-Holmsen: A pioneer in Svalbard. *Polar Research*, *25*(1), 1–13. https://doi.org/10.1111/j.1751-8369.2006.tb00146.x

Fyfe, G. (2006). Sociology and the social aspects of museums. *A Companion to Museum Studies*, 33–49.

Groom, Q. J., 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

Groom, Q. J., O'Reilly, C., & Humphrey, T. (2014). Herbarium specimens reveal the exchange network of british and irish botanists, 1856–1932. *New Journal of Botany*, *4*(2), 95–103. https://doi.org/10.1179/2042349714Y.0000000041

Hipp, R. D. (2020). *SQLite* (Version 3.31.1). https://www.sqlite.org/index.html

Hu, Y. (2011). Algorithms for visualizing large networks. *Combinatorial Scientific Computing*, *5*(3), 180–186.

Lindon, H. L., Gardiner, L. M., Brady, A., & Vorontsova, M. S. (2015). Fewer than three percent of land plant species named by women: Author gender over 260 years. *TAXON*, *64*(2), 209–215. https://doi.org/https://doi.org/10.12705/642.4

Lourenço, M. C., & Dias, J. P. S. (2017). "Time capsules" of science: Museums, collections, and scientific heritage in portugal. *Isis*, *108*(2), 390–398. https://doi.org/10.1086/692690

Mignan, A. (2018). *Metacollecting or the process of collecting collections, with examples from the tricottet collection colligo, 1*. https://perma.cc/YQ5W-BN5Z

Resvoll-Holmsen, H. (1932). Om betydningen av det uensartede i våre skoger. *Tidsskrift for Skogbruk*, *40*, 270–275.

Sastre, C. (2003). Alicia Lourteig (1913-2003). *Adansonia*, *25*(2), 149–150. https://sciencepress.mnhn.fr/en/periodiques/adansonia/25/2/alicia-lourteig-1913-2003

Siracusa, P. C. de, Gadelha, L. M., & Ziviani, A. (2020). New perspectives on analysing data from biological collections based on social network analytics. *Scientific Reports*, *10*(1), 1–10. https://doi.org/10.1038/s41598-020-60134-y