Datathons: fostering equitability in data reuse in ecology

Authors: Datathon 2022 Consortium

Corresponding author(s): *Stephanie Jurburg* (s.d.jurburg@gmail.com), *Maria J. Álvarez Blanco* (maria.alvbla@gmail.com)

Datathon 2022 Consortium

Stephanie Jurburg [1]*, María J. Álvarez Blanco [1,2]*, Antonis Chatzinotas [1,3,4], Anahita Kazem [2,5] , Birgitta König-Ries [2,5], Doreen Babin [6], Kornelia Smalla [6], Victoria Cerecetto [6,7],Gabriela Fernandez-Gnecco [6,8], Fernanda Covacevich [8,9], Emilce Viruel [10], Yesica Bernaschina [11], Carolina Leoni [7,11], Silvia Garaycochea [7,12], Jose. A Terra [13], Pablo Fresia Coronel [14], Eva Lucía Margarita Figuerola [15,16], Luis Gabriel Wall [15,17], Julieta Mariana Covelli [17], Ana Carolina Agnello [18],Esteban Emanuel Nieto [18], Sabrina Festa [18], Lina Edith Dominici [19,] Marco Allegrini [20], María Celina Zabaloy [20], Marianela Estefanía Morales [20,21], Leonardo Erijman [23,23], Anahi Coniglio [24], Fabricio Dario Cassán [24], Sofia Nievas [24], Diego M. Roldán [25,26], Rodolfo Menes [26,27], Patricia Vaz Jauri [28,29], Carla Silva Marrero [28], Adriana Montañez Massa [28], María Adelina Morel Revetria [28,30], Ana Fernández-Scavino [31], Luciana Pereira Mora [31], Soledad Martínez [32], Juan Pablo Frene [33,34]

1. Department of Environmental Microbiology, Helmholtz Centre for Environmental Research (UFZ), 04318 Leipzig, Germany

2. Data and Code Unit (iBID), German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Saxony, Germany.

3. German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany

4. Institute of Biology, Leipzig University, 04103 Leipzig, Germany

5. Department of Mathematics and Computer Science, Friedrich Schiller University Jena, 07743 Jena, Thüringen, Germany

6. Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Epidemiology and Pathogen Diagnostics, 38104 Braunschweig, Germany

7. Instituto Nacional de Investigación Agropecuaria (INIA), Área de Recursos Naturales, Producción y Ambiente, Estación Experimental INIA Las Brujas, Ruta 48 km 10, Canelones, Uruguay.

8. Instituto de Investigaciones en Biodiversidad y Biotecnología-Consejo Nacional de Investigaciones Científicas y Técnicas (INBIOTEC-CONICET), Mar del Plata, Buenos Aires, Argentina

9. Instituto Nacional de Tecnología Agropecuaria, Estación Experimental Agropecuaria Balcarce (INTA, EEA Balcarce), Balcarce, Buenos Aires, Argentina.

10. Instituto de Investigación Animal del Chaco Semiárido (IIACS), Centro de Investigaciones Agropecuarias (CIAP), Instituto Nacional de Tecnología Agropecuaria (INTA), Tucumán, Argentina

11. Instituto Nacional de Investigación Agropecuaria (INIA), Sistema Vegetal Intensivo, Estación Experimental INIA Las Brujas, Ruta 48 km 10, Canelones, Uruguay.

12. Instituto Nacional de Investigación Agropecuaria (INIA), Área Mejoramiento Genético y Biotecnología Vegetal, Estación Experimental INIA Las Brujas, Ruta 48 km 10, Canelones, Uruguay.

13. Instituto Nacional de Investigación Agropecuaria (INIA), Sistema Arroz-Ganadería, Estación Experimental INIA Treinta y Tres, Ruta 8 km 282, Treinta y Tres, Uruguay.

14. Unidad Mixta Pasteur + INIA (UMPI), Institut Pasteur de Montevideo, Montevideo, Uruguay

15. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina.

16. Instituto de Biociencias, Biotecnología y Biología Traslacional, Departamento de Fisiología y Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina.

17. Laboratorio de Bioquímica y Biología de Suelos, Centro de Bioquímica y Microbiología de Suelos, Universidad Nacional de Quilmes (UNQ), Bernal, Buenos Aires, Argentina.

18. Centro de Investigación y Desarrollo en Fermentaciones Industriales (CINDEFI, CONICET-UNLP), La Plata, Argentina

19. Centro de Investigación y Desarrollo en Tecnología de Pinturas y Recubrimientos (CIDEPINT, CICPBA-CONICET-UNLP), La Plata, Argentina.

20. Centro de Recursos Naturales Renovables de la Zona Semiárida (CERZOS, CONICET-UNS), Bahía Blanca, Buenos Aires, Argentina

21. Departamento de Agronomía, Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina

22. Instituto de Investigaciones en Ingeniería Genética y Biología Molecular "Dr Héctor N Torres" (INGEBI, CONICET-UBA), Buenos Aires, Argentina

23. Departamento de Fisiología, Biología Molecular y Celular "Dr Héctor Maldonado", Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA).

24. Laboratorio de Fisiología Vegetal y de la Interacción Planta Microorganismo (LFVIPM), Instituto de Investigaciones Agrobiotecnológicas (INIAB-CONICET), Facultad de Ciencias Exactas Físico-Químicas y Naturales, Universidad Nacional de Río Cuarto (UNRC), Río Cuarto, Córdoba, Argentina.

25. Departamento de Bioquímica y Genómica Microbianas, Instituto de Investigaciones Biológicas Clemente Estable (IIBCE), Ministerio de Educación y Cultura, Montevideo, Montevideo, Uruguay.

26. Laboratorio de Ecología Microbiana Medioambiental, Microbiología, Facultad de Química y Unidad Asociada del Instituto de Química Biológica, Facultad de Ciencias, Universidad de la República Uruguay (UdelaR), Montevideo, Montevideo, Uruguay.

27. Laboratorio de Microbiología, Unidad Asociada del Instituto de Química Biológica, Facultad de Ciencias, Universidad de la República (UdelaR), Montevideo, Montevideo, Uruguay.

28. Laboratorio de Microbiología de Suelos, Instituto de Ecología y Ciencias Ambientales, Facultad de Ciencias, Universidad de la República Uruguay (UdelaR), Montevideo, Montevideo, Uruguay.

29. Laboratorio de Interacción Planta-Microorganismo, Departamento de Bioquímica y Genómica Microbianas, Instituto de Investigaciones Biológicas Clemente Estable (IIBCE), Montevideo, Montevideo, Uruguay.

30. Laboratorio de Microbiología Molecular, Departamento de Bioquímica y Genómica Microbianas, Instituto de Investigaciones Biológicas Clemente Estable (IIBCE), Montevideo, Montevideo, Uruguay.

31. Área Microbiología, Departamento de Biociencias, Facultad de Química, Universidad de la República Uruguay (UdelaR), Montevideo, Montevideo, Uruguay.

32. Laboratorio de Biotecnología, Departamento de Biociencias, Unidad de Análisis de Agua, Facultad de Química, Universidad de la República Uruguay (UdelaR), Montevideo, Montevideo, Uruguay.

33. Future Food Beacon of Excellence, University of Nottingham, LE12 5RD, United Kingdom

34. School of Biosciences, University of Nottingham, LE12 5RD, United Kingdom

Emails and ORCID

Stephanie Jurburg (s.d.jurburg@gmail.com, ORCID 0000-0002-7701-6030)

María J. Álvarez Blanco (maria.alvbla@gmail.com, ORCID 0009-0001-5930-9606)

Antonis Chatzinotas (antonis.chatzinotas@ufz.de, ORCID 0000-0002-0387-9802)

Anahita Kazem (anahita.kazem@idiv.de, ORCID 0000-0003-2475-132X)

Birgitta König-Ries (birgitta.koenig-ries@uni-jena.de, ORCID 0000-0002-2382-9722)

Doreen Babin (doreen.babin@julius-kuehn.de, ORCID 0000-0001-7144-8898)

Kornelia Smalla (kornelia.smalla@julius-kuehn.de, ORCID 0000-0001-7653-5560)

Victoria Cerecetto (victoria.cerecetto@julius-kuehn.de/vcerecetto@inia.org.uy, ORCID 0000-0001-6180-6153)

Gabriela Fernandez-Gnecco (gabyfernandezgnecco@gmail.com, ORCID 0000-0002-1857-1237)

Fernanda Covacevich (fercovacevich@gmail.com, ORCID 0000-0002-6919-6422)

Emilce Viruel (emiviruel@gmail.com, ORCID 0000-0002-7453-4297)

Yesica Bernaschina (ybernaschina@inia.org.uy, ORCID 0000-0002-8296-0456)

Carolina Leoni (cleoni@inia.org.uy, ORCID 0000-0002-3981-564X)

Silvia Garaycochea (sgaraycochea@gmail.com, ORCID 0000-0003-1629-0746)

Jose. A Terra (jterra@inia.org.uy, ORCID 0000-0003-0713-9788)

Pablo Fresia Coronel (pfresia@pasteur.edu.uy)

Eva Lucía Margarita Figuerola (efiguerola@fbmc.fcen.uba.ar, ORCID 0000-0002-7784-1452)

Luis Gabriel Wall (wall.luisgabriel@gmail.com, ORCID 0000-0003-0880-8023)

Julieta Mariana Covelli (julieta.covelli@unq.edu.ar)

Ana Carolina Agnello (aagnello@biotec.quimica.unlp.edu.ar, ORCID 0000-0003-3490-1416)

Esteban Emanuel Nieto (estebanenieto@gmail.com, ORCID 0000-0002-9161-4993)

Sabrina Festa (sfesta@biotec.quimica.unlp.edu.ar, ORCID 0000-0002-3683-9305)

Lina Edith Dominici (l.dominici@cidepint.ing.unlp.edu.ar, ORCID 0009-0001-2706-8182)

Marco Allegrini (mallegrini@cerzos-conicet.gob.ar, ORCID 0000-0002-3597-0387)

María Celina Zabaloy (mzabaloy@uns.edu.ar, ORCID 0000-0003-1155-292X)

Marianela Estefanía Morales (mmorales@cerzos-conicet.gob.ar, ORCID 0000-0002-2721-1428)

Leonardo Erijman (erijman@dna.uba.ar, ORCID 0000-0001-7525-2112)

Anahi Coniglio (nconiglio@exa.unrc.edu.ar)

Fabricio Dario Cassán (fcassan@exa.unrc.edu.ar, ORCID 0000-0002-9776-0262)

Sofia Nievas (snievas@exa.unrc.edu.ar)

Diego M. Roldán (droldan@fcien.edu.uy, ORCID 0000-0002-8900-9815)

Rodolfo Menes (jmenes@gmail.com, ORCID 0000-0001-8968-6149)

Patricia Vaz Jauri (pvazjauri@fcien.edu.uy, ORCID 0009-0001-7965-5901)

Carla Silva Marrero (csilva@fcien.edu.uy)

Adriana Montañez Massa (montanez.massa@gmail.com, ORCID 0000-0001-6516-9008)

María Adelina Morel Revetria (mmorel@fcien.edu.uy, ORCID 0000-0002-9064-5675)

Ana Fernández-Scavino (afernand@fq.edu.uy, ORCID 0000-0001-8484-2392)

Luciana Pereira Mora (lupereira@fq.edu.uy, ORCID 0000-0003-1233-6348)

Soledad Martínez (msmartinez@fq.edu.uy, ORCID 0009-0000-5023-871X)

Juan Pablo Frene (juanpifrene@hotmail.com, ORCID 0000-0001-7754-1591)

Abstract

Approaches to rapidly collect global biodiversity data are increasingly important, but biodiversity blindspots persist. We organized a three-day Datathon event to improve the openness of local biodiversity data, and facilitate data reuse by local researchers. The first Datathon, organized among microbial ecologists in Uruguay and Argentina assembled the largest microbiome dataset in the region to date, and formed collaborative consortia for microbiome data synthesis.

Main text

Like other branches of life, the global microbiome is under threat [1], and documenting the world's microbial diversity is more urgent than ever before. A global coverage of biodiversity data is essential for developing in-depth ecological knowledge of microbial systems [2] and harnessing them as sources of biotechnological innovation [3]. As DNA sequencing technologies have greatly advanced, cataloging the world's microbiomes is now feasible. Over the past decade, sequencing-based assessments of bacterial diversity (i.e., metabarcoding or amplicon sequencing) have grown exponentially [4], but global blind spots in reusable microbiome data persist [5], often affecting the regions that are predicted to undergo the greatest rates of anthropogenic change, and therefore the greatest biodiversity loss [1].

Sequencing-based biodiversity assessments are necessary for most microbiomes, which cannot be characterized through conventional observation, but require substantial financial investments. Several studies have reported a disproportionately higher availability of microbiome data from wealthier countries [5,6]. For example, a systematic literature review of global soil biodiversity research (much of which relies on sequencing), found that only 8% of the studies surveyed originated from Latin America and Africa [7]. Improving data archiving practices is a cost-effective first step towards improving the coverage of compiled, global microbiome data, but requires explicit consideration of the associated costs and benefits, especially to the researchers producing the data.

In ecology, research is disproportionately performed by researchers from high income countries [2], and while data collected from biodiversity blind spots are necessary for global syntheses, synthesis research is seldom performed by scientists from the poorly represented regions, who receive little direct benefit from making their data available. While data citations allow data creators to receive credit for their work, they do not encourage equitable participation in data reanalyses. Decentralizing microbiome data reuse can close the geographic gap between data producers and reusers, improving ecological research [8] in numerous ways. First, data reuse allows scientists to produce high-quality research regardless of their access to infrastructure or funding. Second, the prospect of reusing data may serve as an incentive to archive it publicly, increasing the amount and quality of available, reusable microbiome data in countries with limited research funding. Third, as global participation in synthetic microbiome research increases, so should the diversity of perspectives in the field [9]. Finally, a greater global participation in data reuse may reduce language barriers in synthetic research, which are pervasive [10].

To improve equitability in microbiome synthesis science, it is essential to acknowledge available infrastructures and their limitations, provide educational support and training, build collaborative networks, and credit collaborators [2,9]. We organized a binational data collection and reuse event (*Datathon*) in Argentina and Uruguay, two countries which are poorly represented in global microbiome syntheses [5,6,11]. For example, the Earth Microbiome Project dataset [11] contains only nine microbiome samples from Argentina, and none from Uruguay.

The Datathon provided support to microbial ecologists in archiving and reusing metabarcoding sequence data and brought researchers together to create a common microbiome dataset for Argentina and Uruguay. By centralizing available raw data and including related bibliographic, technical, and experimental materials in a single online database, we aimed to improve the discoverability and reusability of microbial sequence data in the region, while giving data producers academic credit for their work and creating a valuable resource to foster research in a biodiversity blind spot. Crucially, we aimed to stimulate synthesis research by Datathon participants using the newly deposited sequences.

The Datathon was organized over the course of three days in October 2022 in a hybrid format, and each day focused on a different, interconnected aim (Figure 1). The event was preceded by a week-long, intensive statistics and bioinformatics course, which engaged early career researchers. On the Datathon's first day (themed *Inspire*), we held a hybrid symposium focused on the history of synthetic research in ecology, the relevance of 'Open Science' in biodiversity research, and the potential and outcomes of recent global biodiversity data syntheses, available in the Datathon website (https://micoda.idiv.de/datathons.jsp).

The second day (themed *Support*) focused on hands-on sequence data and metadata deposition, and was held remotely to allow all participants access to their work space. Custom, online step-by-step guides were developed to support sequence data preservation and publication process to NCBI's Sequence Read Archives in English and Spanish (freely available in Spanish https://github.com/MariaAlvBla/Dataton-2022/wiki and English https://github.com/MariaAlvBla/NCBI-Tutorial/wiki), in line with the FAIR Principles [12]. Guides contained a custom metadata sheet, which included specific fields to allow for the rapid integration of all datasets and reanalysis following the Datathon, and to give greater visibility to the original publications of the data creators. In addition to standard NCBI metadata fields, we included fields for technical information (e.g., DNA extraction method and sample amount), and the DOI for any scientific article accompanying the initial publication of the sequence data. In addition, given the broad range of environments sampled within the context of microbiome research, we developed a three-level ontology based on the Earth Microbiome Project [11], where users could select their sample's realm (e.g., aquatic, mineral, host-associated), broad-scale environment (e.g., soil, freshwater), and complete one description per sample (e.g. "agricultural soil from soy farm"). If data had been made publicly available prior to the Datathon, participants could complete the metadata sheet with the original accession numbers to expedite data integration. All members were invited to join an online, dedicated *Slack* group, which included a help desk channel, staffed by organizers who had previously tested the step-by-step guides. The help desk allowed participants to obtain more personalized help when needed, allowed organizers to clarify common questions that emerged during the event (e.g., "what is the most appropriate unit for a sample size?") and improve available materials for events organized in the future, and resulted in participants helping each other and networking. improving the quality of the deposited datasets and contributing towards community-building.

The third day (themed *Collaborate*) focused on harnessing the deposited data for reuse by participants and on developing collaborative networks. Following a general summary of the collected data, participants created *Slack* channels to brainstorm and develop projects to reuse the data, and secure funding to pursue these ideas. Then, collaborative synthesis projects were voted on, and leaders were collectively selected for each. After the meeting, all participants received a detailed summary of the data resource created and the planned synthesis projects and activities, and could opt-in to each, allowing participants to contribute towards the organization of the next Datathon.

In total, 59 scientists, from various academic levels including PhD-level academics (54%), PhD students (24%), and MSc and BSc scientists and students (22%) participated in the Datathon. Of these, 52% participated in all days and archived sequence data, 32% participated in the symposium and training sessions, and 15% joined only one of the sessions.

Except for one case, all datasets were deposited by PhD-level academics or PhD students, suggesting that other student participants and those with MSc and BSc degrees benefitted primarily from the *Inspire* and *Support* components of the Datathon.

The event archived 913 samples from 22 projects in NCBI's Sequence Read Archives (Figure 1), greatly enriching microbiome data from the region (the previous, largest survey of Argentinean pampas included 126 samples, for example [13]). Deposition to NCBI ensures that the data remains publicly accessible in the long-term, and facilitates integration with other publicly available datasets (NCBI accession numbers and metadata, are available in the Datathon website). Of the contributed projects, 55% (33% of samples) were previously unarchived. Previously unarchived projects ranged in size between 4 and 129 samples, and 64% of the samples were deposited as projects with 75 samples or less, illustrating the potential of Datathon events in promoting the archiving of small datasets. Furthermore, the custom data deposition guides developed for the event remain publicly available as living educational documents for users aiming to deposit microbiome sequence data in the future, and serve as a model for translation into other languages. The compiled dataset is dominated by soil microbiome samples, likely because the initiative began outreach through the Soil BON [14] network of researchers, illustrating the influence of the networking approach.

Notably, the activities proposed during the *Collaborate* component have resulted in the formation of a close-knit scientific community. One year after the first Datathon, 10 of the original participants and organizers held a second, larger Datathon encompassing all of Latin America. Members who deposited data continued to regularly exchange ideas, technical information, and funding opportunities in a group mailing list. Over the coming year, this consortium will collaborate with local researchers to organize a similar event in Africa, and will coordinate a series of hybrid bioinformatics and statistics courses designed to support researchers as they continue to develop their synthesis projects.

Equitable participation from researchers globally through synthesis work can reduce disparities arising from differential access to funding, in turn reducing existing biases in research [6], increasing the scope/breadth of research and bolstering transparency in the receipt of academic credit [15], and excellence in science. The rapid, coordinated public archiving of microbiome sequence data from 913 samples within three days demonstrates the tremendous potential of equitable data consolidation approaches to shed light on biodiversity blind spots, in both microbiome research and other areas of ecology. The future reuse of these data by researchers from the region that produced the data will likely advance the collective scientific knowledge of the microbiomes of South America, how they are affected by local anthropogenic change, and how local policies may mitigate microbial diversity loss.

Declaration of interests

The authors declare no competing interests

Acknowledgements

We thank our symposium speakers, R. van Klink, E. Ladouceur, S. Blowes, and C. Guerra. This project was funded by the Flexpool program of the German Centre for Integrative Biodiversity Research (iDiv) – Halle, Jena, Leipzig (346001300-22). We acknowledge the support of iDiv funded by the German Research Foundation (DFG– FZT 118, 202548816).

Bibliography

1 Averill, C. *et al.* (2022) Defending Earth's terrestrial microbiome. *Nat. Microbiol.* 7, 1717–1725

2 Nuñez, M.A. *et al.* (2021) Making ecology really global. *Trends Ecol. Evol.* 36, 766–769

3 Vuong, P. *et al.* (2022) The little things that matter: how bioprospecting microbial biodiversity can build towards the realization of United Nations Sustainable Development Goals. *npj biodivers.* 1, 4

4 Ahmed, S.A.J.A. *et al.* (2022) Large scale text mining for deriving useful insights: A case study focused on microbiome. *Front. Physiol.* 13, 933069

5 Guerra, C.A. *et al.* (2020) Blind spots in global soil biodiversity and ecosystem function research. *Nat. Commun.* 11, 3870

6 Abdill, R.J. *et al.* (2022) Public human microbiome data are dominated by highly developed countries. *PLoS Biol.* 20, e3001536

7 El Mujtar, V. *et al.* (2019) Role and management of soil biodiversity for food security and nutrition; where do we stand? *Glob. Food Sec.* 20, 132–144

8 Aubin, I. *et al.* (2020) Managing data locally to answer questions globally: The role of collaborative science in ecology. *J. Veg. Sci.* 31, 509–517

9 Oduaran, O.H. and Bhatt, A.S. (2022) Equitable partnerships and the path to inclusive, innovative and impactful human microbiome research. *Nat. Rev. Gastroenterol. Hepatol.* 19, 683–684

10 Konno, K. *et al.* (2020) Ignoring non-English-language studies may bias ecological meta-analyses. *Ecol. Evol.* 10, 6373–6384

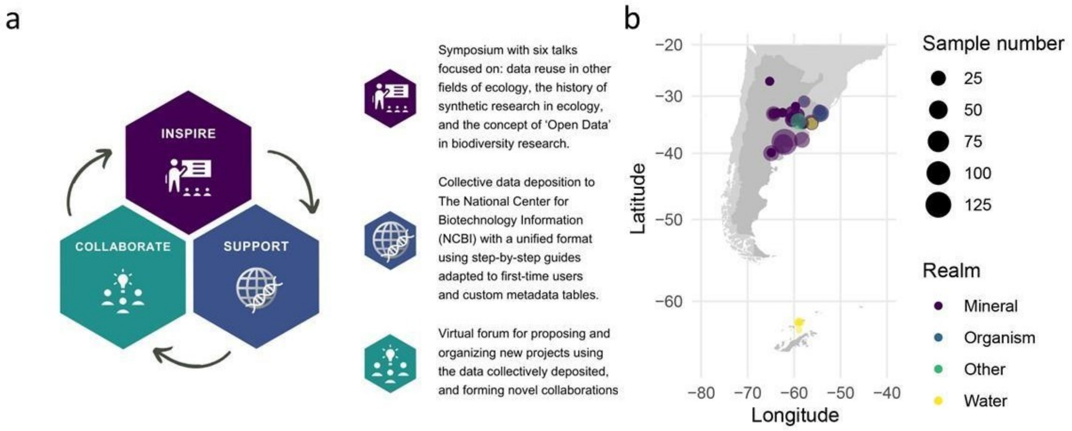
11 Thompson, L.R. *et al.* (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463

12 Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018

13 Rascovan, N. *et al.* (2013) The PAMPA datasets: a metagenomic survey of microbial communities in Argentinean pampean soils. *Microbiome* 1, 21

14 Guerra, C.A. *et al.* (2021) Tracking, targeting, and conserving soil biodiversity. *Science* 371, 239–241

15 Eichhorn, M.P. *et al.* (2020) Steps towards decolonising biogeography. *Front. Biogeogr.* 12,



Symposium with six talks focused on: data reuse in other fields of ecology, the history of synthetic research in ecology, and the concept of 'Open Data' in biodiversity research.

Collective data deposition to The National Center for Biotechnology Information (NCBI) with a unified format using step-by-step guides adapted to first-time users and custom metadata tables.

Virtual forum for proposing and organizing new projects using the data collectively deposited, and forming novel collaborations

Figure 1. Equitable data archiving and reuse practices must consider the needs of participants. a. In designing the Datathon, we considered the need for background information on the potential of data synthesis, providing support during data deposition, and the establishment of collaborative networks, to be of primary importance in furthering microbiome synthesis research. b. The project resulted in the deposition and consolidation of 913 samples, and a much-improved coverage of the Argentina-Uruguay region.