



Soil biodiversity joint data mobilization call

Final Report
(February 2023)



1. Structure of the call

This call emerged as a joint pilot project between the Global Soil Biodiversity Initiative, in the context of SoilBON, and the Global Biodiversity Information Facility *with the aim of mobilizing soil biodiversity datasets, but also to create the beginning of a community of data holders of soil taxa datasets*. This call solicited applications by researchers/research groups from the group of SoilBON collaborators, for projects and processes related to digital data mobilization of already collected soil biodiversity data. Soil biodiversity data holders not currently affiliated with SoilBON had the option to join SoilBON as data providers and integrate the pilot in that way. Applications (researchers and data) coming from underrepresented regions were given preference as the call also aimed to have a global coverage. This initiative aimed to substantially increase the amount of species distribution data openly available at GBIF and, with that, improve the capacity of soil ecologists around the world to address key ecological questions. All data submitted through this initiative is to follow the data standards supported by GBIF, and published through GBIF. Attribution of the data, as well as the credits for the dataset remain with the data provider.

For this pilot 25.000 EUROS were made available with the initial goal of funding between 5 and 7 projects with up to \$5000 USD per project. This funding was made available through the collaboration between SoilBON and the Global Biodiversity Information Facility. All information about the call and was made available through the GSBI website (here: <https://www.globalsoilbiodiversity.org/soilbon-call-2?rq=gbif>). As a result of the call made in 2022, nineteen different projects applied covering a wide range of institutions, career stages of the applicants and also geographies. The table below provides a synthesis of the projects that have applied to the call.

Table 1 List of applicants to the call in September 2022

PI	Institution	Country	Biome(s)	Continent(s)	No. datasets	Level identified	Taxa	Dataset classes
S. Geisen	Laboratory of Nematology, Wageningen University	Netherlands	Desert; Grassland; Savannah; Taiga; Temperate Forest; Tropical Rainforest; Tundra	Africa; Antarctica; Asia; Australia/Oceania; Europe; South America; North America	2	Genus, family, trophic group	Nematodes	Sampling Event; Occurrence; Checklist
M. Bahram	Swedish University of Agricultural Sciences	Sweden	Desert; Freshwater; Grassland; Marine; Savannah; Taiga; Temperate Forest; Tropical Rainforest; Tundra	Africa; Antarctica; Asia; Australia/Oceania; Europe; South America; North America	1	Operational Taxonomic Unit	Archaea; Bacteria; Fungi; Nematodes; Protista	Sampling Event; Occurrence; Checklist

C. P. Peña-Venegas	Amazonic Inst. for Scien. Research Sinchi	Colombia	Tropical Rainforest	South America	1	Genus; species	Ants (Formicidae); Termites (Isoptera)	Sampling Event
K. E. Kwadjo	University NANGUI ABROGOUA	Cote d'Ivoire	Savannah; Tropical Rainforest	Africa	14	Species; Genus; Family	Ants; Bacteria; Beetles; Collembola; Diplura; Earthworms; Enchtraeids; Fungi; Isopods; Mites (Acari); Nematodes; Pseudoscorpiones; Termites; Arachnids	Occurrence; Checklist
G. Brown	Embrapa Forestry/UFPR	Brazil	Grassland; Savannah; Tropical Rainforest	Africa South America; North America	3	Order, class or family	Ants; Beetles; Diplura; Earthworms; Enchtraeids; Isopods; Pseudoscorpiones; Termites; others	Sampling Event
J. V. Lemos Cavalcante de Oliveira	Universidade Estadual da Paraiba	Brazil	Tropical Rainforest	South America	5000	Species	Collembola	Occurrence
O. Marushchak	NGO "Ukrainian Nature Conservation Group"	Ukraine	Grassland; Temperate Forest	Europe	3	Species	Mites (Acari)	Occurrence
M. Pinto	Charles Darwin Foundation	Ecuador	Desert; Tropical Rainforest	South America	2	Species	Ants (Formicidae)	Occurrence; Checklist
G. Soliman	Permaculture Association Britain	United Kingdom	Desert	South America	23	Genus	Archaea; Bacteria; Earthworms; Fungi; Protista	Sampling Event; Occurrence
K. P. Villanuev	Mindanao State University	Phillipines	Marine	Asia	100	Species; Genus	Bacteria	Sampling Event
N. Ascarrunz	Instituto Boliviano de Investigacion Forestal	Bolivia	Grassland	South America	3	Family; Genus	Bacteria; Fungi	Occurrence
A. Góes-Neto	Federal University of Minas Gerais	Brazil	Tropical Rainforest	South America	1	Genus	Bacteria	Occurrence
C. Ramírez-Pérez	Universidad Autónoma de Santo Domingo, Instituto de Investigaciones Botánicas y Zoológicas	Dominican Republic	Tropical Rainforest	North America	2	order; family	Beetles (Coleoptera); Hymenoptera; Diptera; Hemiptera	Occurrence
P. DAS	CVIJAYGARH JYOTISH RAY COLLEGE	India	Marine; Tropical Rainforest	Asia	20	Species and/or genus	Algae; Archaea; Bacteria; Fungi	Sampling Event; Occurrence
M. Logachev	Universidad de Granada	Spain	Freshwater; Grassland; Temperate Forest	Europe	2	Species	Collembola	Occurrence; Checklist

C. Duarte Ritter	Universidade Federal do Parana	Brazil	Tropical Rainforest	South America	36	Family, genus, species	Bacteria; Fungi; Protista	Sampling Event
M. Zafar	Quaid-i-Azam University Islamabad	Pakistan	Grassland	Asia	26	Species; genus	Soil Organic Matters	Sampling event; Occurrence; Checklist
D. Guasconi	Dept of Physical Geography, Stockholm University	Sweden	Grassland	Europe	2	Species; Genus; Functional traits	Bacteria; Fungi; plants	Sampling Event
J. Nantongo	National Agricultural Research Organization	Uganda	Tropical Rainforest	Africa	2	Genus; species	Bacteria; Fungi	Sampling Event; Occurrence

1.1. Funded projects and expected results

Of the 19 projects that were submitted, 5 were finally funded. These refer to the projects submitted by the Laboratory of Nematology of the Wageningen University in the Netherlands, the Swedish University of Agricultural Sciences, the Amazonic Institute for Scientific Research Sinchi from Colombia, the University of Nangui Abrogoua in Cote d'Ivoire, and the Embrapa Forestry/UFPR from Brazil.

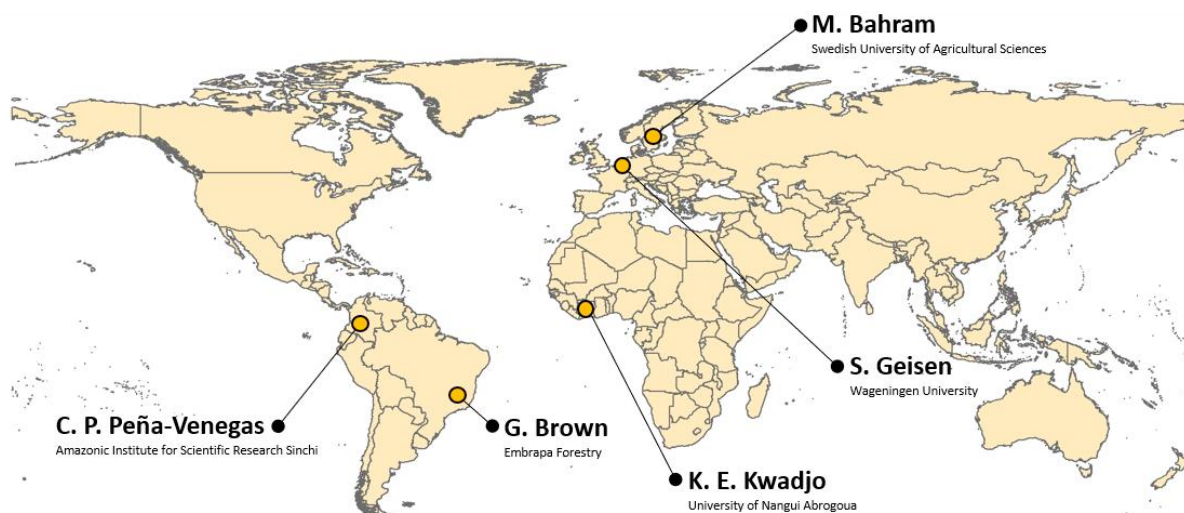


Figure 1 Global distribution of the funded projects

The funded projects expected to deliver 21 datasets that cover a diversity of taxonomic groups, including Archaea; Bacteria; Fungi, Enchtraeids, Isopods, Mites, Nematodes; Protists, Ants; Termites, Beetles (Coleoptera), Collembola, Diplura, and Earthworms. Of these, three of the projects cover cross-continental sampling locations and two are focused on a national (Colombia) or regional extent (Sub-Saharan Africa).

2. Results from the call

After the project end, in February 2023, 42 datasets were published covering all the groups that were initially expected to be covered with a total of 7.242.214 occurrences. Of these occurrences, 7.064.657 came from a single dataset with the remaining datasets totaling 177.557 occurrences. The resulting datasets are described in the table below. Out of these, there is one dataset that did not comply with the original purpose of the call by the date of the project end, since it has an embargo associated to it due to publication efforts of the dataset. This refers to a European dataset on soil Nematodes. That said, all the applicants saw this effort as a starting point with many of them pledging increasing their contributions in the coming months. The datasets published by the Brazilian applicant are not fully published within GBIF but rather in the Brazilian node of GBIF. This is due to an issue that is raised by the applicant (see next section for more details).

Table 2 Datasets published in the scope of the pilot call on soil biodiversity datasets

PI	Dataset	Taxa	Occurrences
C. P. Peña-Venegas	https://www.gbif.org/dataset/b921c99a-c116-4ac6-8271-c246542117c3	Ants and Termites	6242
	https://doi.org/10.15468/pcunej	Isopods	79
K. E. Kwadjo	https://doi.org/10.15468/jxb9mf	Myriapods	391
	https://doi.org/10.15468/mfjqh7	Bacteria	33
	https://doi.org/10.15468/xxmkjm	Beetles	706
	https://doi.org/10.15468/pm2s4h	Collembola	273
	https://doi.org/10.15468/nt4na6	Earthworms	513
	https://doi.org/10.15468/s6k93x	Mites	813
	https://doi.org/10.15468/dxz7f2	Nematodes	569
	https://doi.org/10.15468/azm5k9	Termites	1918
	https://doi.org/10.15468/7azhpt	Fungi	1293
	https://doi.org/10.15468/46e8ss	Arachnids	1106
	https://doi.org/10.15468/72275b	Ants	1967
S. Geisen	https://www.gbif.org/dataset/ccb62af8-f3f0-47fe-9238-60b0d7eddfcf	Nematodes	35799
	https://ipt.nl/bif.nl/manage/resource?r=european_nematode	Nematodes*	51923
M. Bahram	https://www.gbif.org/dataset/9f0e1ca6-fb08-4c72-9a4a-1e3b7a528c10	Fungi	7064657
G. Brown	https://ipt.sibbr.gov.br/sibbr/resource?r=1_soil_macrofauna_sampling_godoy_2001	Several	315
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=1_soil_macrofauna_sampling_godoy_2003	Several	720
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=soil_macrofauna_londrina_2001_2005	Several	2,520
	https://ipt.sibbr.gov.br/sibbr/resource?r=15_soil_macrofauna_sampling_lerroville	Several	630
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=4_soil_macrofauna_sampling_dionisio_embrapa	Several	1,440
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=12_soil_macrofauna_sampling_embrapasojaorganico	Several	450
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=soil_macrofauna_vanesca	Several	6,750
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=3_soil_macrofauna_sampling_reservaflorestal_2001	Several	765
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=11_soil_macrofauna_sampling_saojeronimo	Several	720
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=8_soil_macrofauna_sampling_cafeara	Several	3,870
	https://ipt.sibbr.gov.br/sibbr/resource?r=soil_macrofauna_jaguapita	Several	20,250
	https://ipt.sibbr.gov.br/sibbr/resource?r=19_soil_macrofauna_sampling_colorado	Several	7,020
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=6_soil_macrofauna_sampling_campomourao_galerani	Several	720
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=16_soil_macrofauna_sampling_lapacontestado	Several	1,800
	https://ipt.sibbr.gov.br/sibbr/resource?r=soil_macrofauna_ururu	Several	2,700
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=20_soil_macrofauna_sampling_vilavelha	Several	2,700
	https://ipt.sibbr.gov.br/sibbr/resource?r=soil_macrofauna_pontagrossa	Several	6,450
	https://ipt.sibbr.gov.br/sibbr/resource?r=17_soil_macrofauna_sampling_embrapaflorestas	Several	360
	https://ipt.sibbr.gov.br/sibbr/manage/resource?r=soil_macrofauna_nita	Several	1,620
	https://ipt.sibbr.gov.br/sibbr/resource?r=soil_macrofauna_curitiba	Several	4,500
https://ipt.sibbr.gov.br/sibbr/manage/resource?r=soil_macrofauna_antonina	Several	1,440	
https://ipt.sibbr.gov.br/sibbr/manage/resource?r=soil_macrofauna_itatinga	Several	6,192	

* need to clarify embargo policies.

While most of the occurrence data comes from European countries, the datasets provided also had important contributions for the sub-Saharan region of Africa, particularly in terms of groups covered, and south-America. While this is true, the fact is that most of the datasets contributed did not come from Europe nor address the European continent.

3. Suggestions from the pilot to GBIF for future calls

Being a pilot project, it was requested to the applicants to describe their main bottlenecks when engaging with this process. These suggestions come from a community that usually is not engaged in such data mobilization activities and may require special consideration.

Main issues faced by the applicants:

2. The first issue that was raised several times was on how to include soil data (soil characteristics usually related to the occurrence data and required by the soil ecology community) together with the occurrence data. A few options were raised, including publishing the datasets in EDAPHOBASE but these do not solve the issue completely in terms of having a central, non-duplicated, repository for soil biodiversity data.
3. Another issue is related to the inclusion of biomass information into the datasets deposited in GBIF, since this normally only allows for a single variable like abundance, while biomass is typically used in soil ecology.
4. The third issue is related to depth. Depth information is quite relevant in soil ecology (e.g., litter, 0-10cm, 10-20 cm or others). Although the most important aspect for a sampling event and occurrence record is the presence or absence of a particular taxon in the monolith (which was guaranteed by inclusion of individual sample data), aspects more related to the biology of the organisms are lost when data is not inserted at the individual depth layers.
5. The short-term length of the projects was also raised as a potential issue also in line with having to learn an entire new procedure and having to deal with other local institutions (GBIF nodes).
6. Having to deal with several, national, nodes of GBIF may be problematic for such a centralized call, as it requires often several explanations/meetings of the purpose and justification of the datasets being published, with some publishers refusing the publication because they were not the owners of the datasets.

Main suggestions:

1. Extend the duration of future projects to 12 months, as some time is needed to get familiar with how GBIF works, go over the review process of national nodes, and there may also be some delay in receiving the funds depending on the country.

2. Using the PlutoF platform (<https://plutof.ut.ee>) for publishing biodiversity data (in particular DNA-derived occurrence data of fungi) to GBIF may facilitate data upload for users with little bioinformatics background. Nevertheless, this does not fully avoid duplication of records.
3. Since the ultimate objective of the data being published is that it be used for various ecological studies, the details needed in the Metadata section should be prioritized and expanded, particularly for unpublished datasets (for which there are no associated papers or publications). A list of minimum metadata to be provided in these cases should be proposed and properly integrated in the GBIF platform.
4. Videos should also be prepared or made available explaining the publication process and the entry of all the different excel data-tables into the GBIF online system, when there is sampling event, occurrence and measurement data (e.g., of soil parameters). Although tutorials are available online, they are not always easy to follow, or do not have all the different options. Tailored videos or instructions more geared towards the work being performed saves time of the grantees. Currently there is only the option to watch a demo video on the webpage. In the future, it would be convenient to get a manual to introduce biodiversity data mapping in the Darwin Core and so on. Ideally, a script with example datasets of data that likely resemble expected biodiversity data, etc. would help.
5. A global publisher for soil biodiversity data would allow for more integration and a stronger publishing community, as we know better what kind of data to expect and how to format and get it in shape. It also allows to have a vetted community of data contributors. Also, it would help overcome the fact that different countries have a different process and templates in use making it difficult to provide help between research teams of different countries. In case the current model is followed, the GBIF national nodes should be identified before-hand by the granting agency and GBIF, and their contact data sent to the grantees in order to facilitate logistics and allow for more rapid information exchanges (e-mails, telephone) between GBIF nodes and the grantees. This will resolve questions more rapidly and facilitate the publishing of the datasets. Since some nodes are more accessible than others, this also needs to be taken into consideration in terms of the actual time needed to perform the different steps in preparation and publication of the data.
6. Metadata information requires the detail of the high-taxonomic level reported. The format only allowed to include one level. If we select species level, genus without specific epithet identified (species) would not be visible in the DOI webpage. And if we reported to genus level, searchers by species won't provide results. We suggest to include different taxonomic levels to avoid miss information during GBIF searchers.

Finally, it is important to note that it is necessary that the researchers that would like to upload information in GBIF, need experience in Darwin Core or similar databases, and their institution registered in GBIF previously. These steps will help the process, otherwise the process is long, and requires external support to do it, leading to stronger limitations (or demotivation) for publication of soil biodiversity information for new research groups. Furthermore, as there are few people who are experts in analyzing downloaded data from GBIF for ecological purposes, a video or course on this topic could be offered, in order to facilitate the work of using data already deposited related to the taxa and increase the motivation of the data holders. This is especially relevant as more and more soil biodiversity-related data becomes published, and others begin to see the possibilities of using these data for various syntheses and other analyses.