User Requirements Analysis Report

Deliverable D1.1

29 June 2022

Authors

Christos Arvanitidis[1], Cristina Huertas Olivares[1], Quentin Groom[2], Elizabeth Bamford[3], Lucia Vaira[3], Sara Montinaro[3], Ana Casino[4], Donat Agosti[5], Boris Barov[6], Kristina Hristova[6], Lyubomir Penev[6]

Author affiliations

1. *LifeWatch ERIC, Seville, Spain*
2. *Meise Botanic Garden, Meise, Belgium*
3. *LifeWatch ERIC, Lecce, Italy*
4. *CETAF, Brussels, Belgium*
5. *Plazi, Bern, Switzerland*
6. *Pensoft Publishers, Sofia, Bulgaria*

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

This project receives funding from the European Union's Horizon 2020 Research and Innovation action under grant agreement No 101007492.

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	User requirements analysis report
Deliverable n°:	D1.1
Nature of the deliverable:	Report
Dissemination level:	Public
WP responsible:	WP1
Lead beneficiary:	LifeWatch ERIC
Citation:	Arvanitidis, C., Huertas Olivares, C., Groom, Q., Bamford, E., Vaira, L., Montinaro, S., Casino, A., Agosti, D., Barov, B., Hristova, K,. Penev, L. (2022). *User requirements analysis report*. Deliverable D1.1 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	12
Actual submission date:	14

Deliverable status:

Version	**Status**	**Date**	**Author(s)**
1.0	Draft	15 June 2022	Christos Arvanitidis and Cristina Huertas Olivares, LifeWatch ERIC
2.0	Review	17June 2022	Quentin Groom
2.1	Review	17 June 2022	Dimitris Koureas
2.2	Review	20 June 2022	Kristina Hristova
2.3	Review	20 June 2022	Lyubomir Penev
2.4	Review	24 June 2022	Joana Pauperio
3.0	Submission	29 June 2022	Christos Arvanitidis and Cristina Huertas Olivares

The content of this deliverable does not necessarily reflect the official opinions of the European Commission or other institutions of the European Union.

Table of Contents

Executive Summary

Background: Work Package 1 of BiCIKL is focused on the "*Coordination and interoperability of infrastructures through harmonisation of community policies, standards and guidelines*". Two of the specific objectives of WP1 are: (a) "Understanding user requirements from the community in terms of the questions they want to address with the data" and (b) "*Coordination of data standards usage between infrastructures, particularly exchange standards and controlled vocabularies*". Therefore, the design, development and submission of a Deliverable which analyses the users requirements, is a natural pre-requisite in order to achieve the afore-mentioned objectives.

Objectives: The sole objective of D1.1 is to analyse the users' needs for types of data, quality of data and modes of access.

Methods: The methods applied were aiming at accommodating all levels of access and modes of access that can suit the user needs. The BiCIKL partners carefully designed an online questionnaire in order to collect the appropriate information from the broader community. Subsequently, this information was converted to data and finally both uni- and multivariate analyses were applied in order to find potential patterns and infer the processes behind the patterns.

Results: A total of N=72 responses were collected from the members of the broader community. Most of the participants who filled in the questionnaire: (a) are males; (b) they are in the most productive age classes (36-65 years old); (c) work in European countries; (d) require *specimen data*, *taxon names* and *images* for their work; (e) use mostly GBIF, BHL and BOLD as data sources; (f) use wikidata, cloud services and their ORCID ID for purposes other than publications; (g) need standardised data. The multivariate statistics applied showed no specific pattern based on the similarities in the researchers' requirements in the types of data, their quality and modes of access they have. No metadata was found to be associated with the above result.

Conclusion: The study has demonstrated that most of the researchers have a clear preference in high quality standardised data, primarily specimen data, taxon names and images, and they mostly use GBIF, BHL and BOLD as sources of data. However, these preferences are randomly distributed among the researchers who replied to the questionnaire nor are there any specific metadata associated with the above pattern. These results also indicate that there are no biases in the sample of the researchers who participated in this study.

List of Abbreviations

BKH	Biodiversity Knowledge Hub
BIO-ENV	Biotic - environmental variables association
DNA	Deoxyribonucleic acid
HPC	High Performance Computer (Cluster)
IMBBC, HCMR	Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research
nMDS	Non-metric multidimensional scaling

1. Introduction

Work Package 1 of BiCIKL is one of the most integrative WPs of BiCIKL since it deals with harmonisation actions on the community policies, standards and guidelines in order to achieve both coordination and interoperability of the infrastructures as holders of various types of information and data. It is an essential task of BiCIKL in order to achieve the bi-directional linking of literature, taxonomic, DNA sequence and occurrence data, provided by the partnering infrastructures (WPs 6-9).

The needs and requirements of the communities in the type, quality and ways of access is very important for the development of the architecture that allows them to discover and address their questions by using all of the above-mentioned types of information and data in combination. In particular, they inform BiCIKL on how to build its architecture in order to harmonise and coordinate the infrastructures in contributing with new methods and workflows to harvest, liberate, link, reuse data from specimens, samples, sequences, taxonomic names and taxonomic literature. Finally, these requirements are important for the design, development and operation of the Biodiversity Knowledge Hub (BKH) in WPs 2 and 3 (Figure 1).

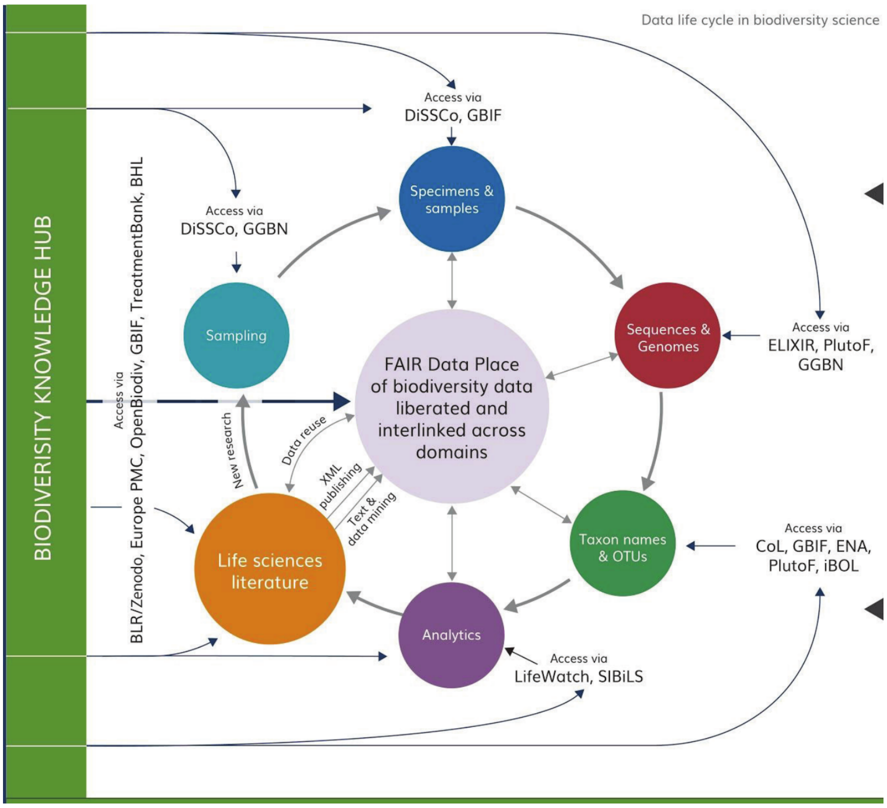


Fig. 1: *Access to types of data and services and the infrastructures of the Consortium involved towards the development and operation of the Biodiversity Knowledge Hub (BKH) (from the BiCIKL proposal text).*

Deliverable 1.1 is conceived like a search for information and analysis on the broad user-base requirements for linked data. These needs must include various aspects such as best practices, types of data that are made accessible, the linkages between them and the modes of access to the above-mentioned categories. Such examples of requirements may include, on one hand, specific requests to single collection entities, such as finding all the specimens and sequences related to a publication. However, the breadth of requests may be vast and may include requests for the whole body of linked works, and further, requests for text and data mining across large corpora of text-based knowledge, with the latter cases to include even research into scientometrics. The aim, therefore, of this deliverable is to accommodate all levels of access and find modes of access that will suit user needs to address their specific scientific questions. Finally, it is expected that these user requirements will foresee future requirements going far beyond the status quo.

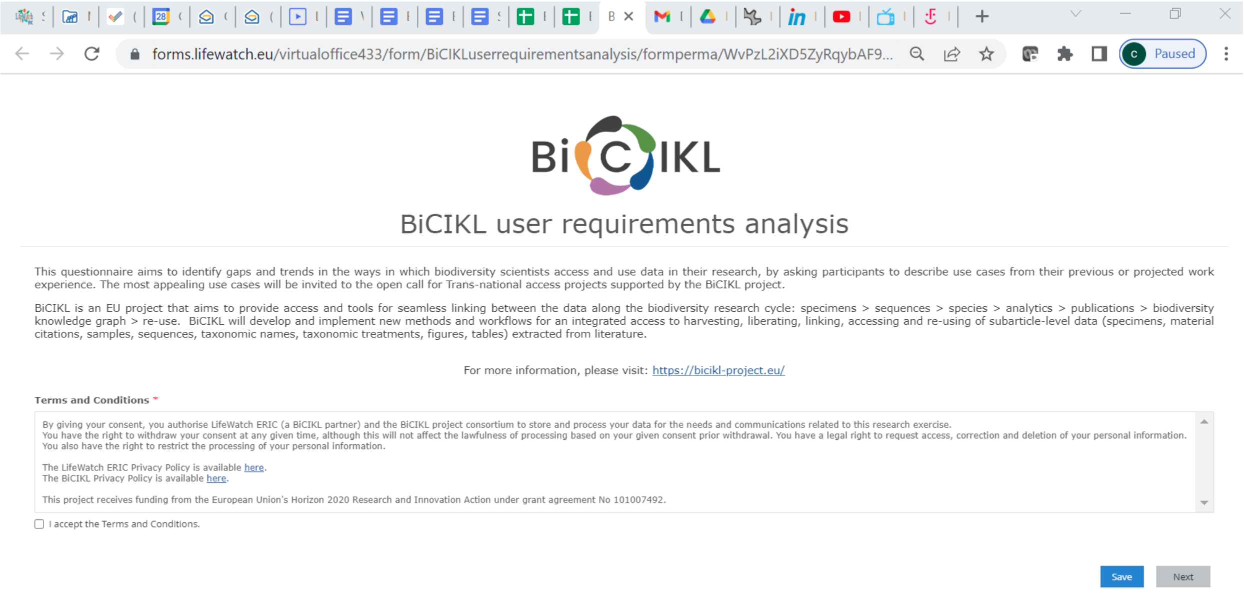
2. Methodology

The discussions between the Consortium partners on how to collect the information and data on the users requirements started early, in month 2 of the project. Representatives of the WPs involved (1-3, 6-9, 11) as well as from the partnering infrastructures participated in this dialogue. Meetings in the context of WP1, Pillar 1, Pillar 2 and Project Management Board, were used to maintain this dialogue and make sure that all different aspects have been taken onboard in the design, development of the questionnaire and the subsequent analysis of the data.

The common understanding recorded during the above meetings was that the main vehicle to collect the information and data on the users requirements and needs should be an online questionnaire. The architecture, components and layout of this questionnaire were also discussed and agreed during this dialogue. The landing (Figure 2) and subsequent pages, the choices researchers had in order to fill in this questionnaire have been exhaustively discussed and agreed among the members of the Consortium. Depending on the scientific questions, the sorts of data and information, the level of quality and means of access, the user needed in order to address her/his question, different options to provide the relevant information were made available. The questionnaire was designed this way in order to collect the maximum amount of information and data to analyse.

The next step was to advertise the questionnaire developed[1] to as many potential user groups of the types of information and data described above as possible. This was achieved by the dissemination of the information to the internal communication channels of the BiCIKL Consortium, the internal channels of each partner and especially those of the infrastructures and finally in the social media and global mailing lists (e.g. taxacom, MARINE-B, etc.).

[1] https://forms.lifewatch.eu/virtualoffice433/form/BiCIKLuserrequirementsanalysis/formperma/WvPzL2iXD5ZyRqybAF960fwfp9Us2bNZ_iJFD0vSnlM


forms.lifewatch.eu/virtualoffice433/form/BiCIKLuserrequirementsanalysis/formperma/WvPzL2iXD5ZyRqybAF9...
Paused
BiCIKL

BiCIKL user requirements analysis

This questionnaire aims to identify gaps and trends in the ways in which biodiversity scientists access and use data in their research, by asking participants to describe use cases from their previous or projected work experience. The most appealing use cases will be invited to the open call for Trans-national access projects supported by the BiCIKL project.

BiCIKL is an EU project that aims to provide access and tools for seamless linking between the data along the biodiversity research cycle: specimens > sequences > species > analytics > publications > biodiversity knowledge graph > re-use. BiCIKL will develop and implement new methods and workflows for an integrated access to harvesting, liberating, linking, accessing and re-using of subarticle-level data (specimens, material citations, samples, sequences, taxonomic names, taxonomic treatments, figures, tables) extracted from literature.

For more information, please visit: https://bicikl-project.eu/

Terms and Conditions *

By giving your consent, you authorise LifeWatch ERIC (a BiCIKL partner) and the BiCIKL project consortium to store and process your data for the needs and communications related to this research exercise.
You have the right to withdraw your consent at any given time, although this will not affect the lawfulness of processing based on your given consent prior withdrawal. You have a legal right to request access, correction and deletion of your personal information.
You also have the right to restrict the processing of your personal information.

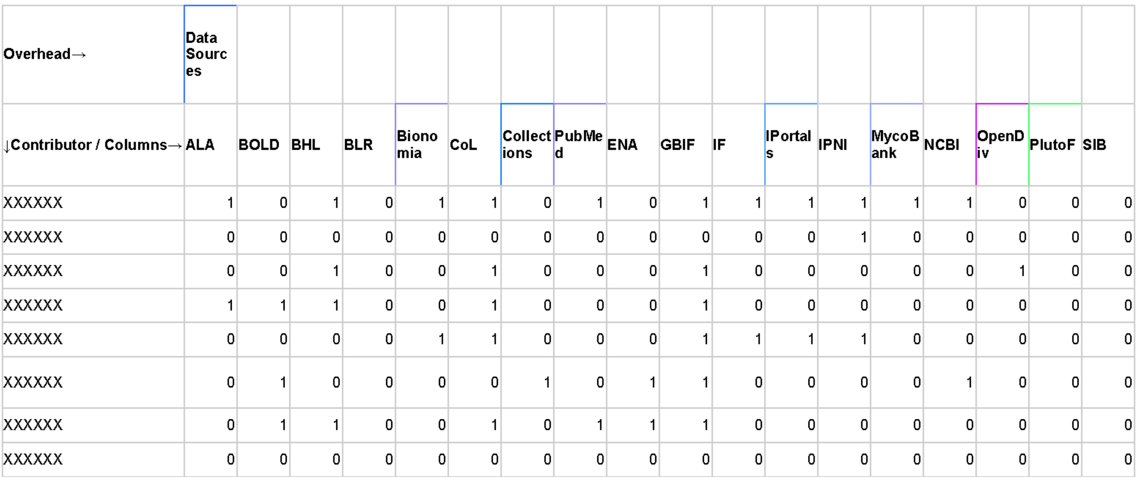
The LifeWatch ERIC Privacy Policy is available here.
The BiCIKL Privacy Policy is available here.

This project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

☐ I accept the Terms and Conditions.

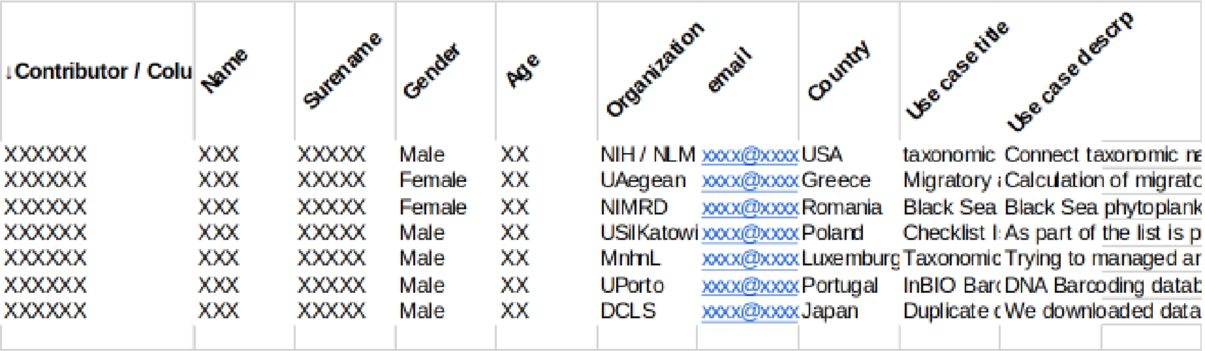
Fig. 2: *Screenshot of the landing page of the BiCIKL user requirements questionnaire.*

The information was collected automatically in a google workbook and then converted to data of numerical values in order to be properly analysed. Two matrices were produced by this process: (a) the primary matrix including the data with binary values of presence, absence (1,0) for all the variables (e.g. types of data, quality and ways of access) (Figure 3); (b) the metadata matrix including all the metadata of the previous data matrix, such as the identity (address ID) of the researchers, their age, sex, affiliation, and other relevant data (Figure 4). This data was also converted to numerical categorical values in order to be used in the analysis. Finally, two methods have been applied to analyse the data: uni- and multivariate statistics.



Overhead→	Data Sources																	
↓Contributor / Columns→	ALA	BOLD	BHL	BLR	Bionomia	CoL	Collections	PubMed	ENA	GBIF	IF	IPortals	IPNI	MycoBank	NCBI	OpenDiv	PlutoF	SIB
XXXXXX	1	0	1	0	1	1	0	1	0	1	1	1	1	1	1	0	0	0
XXXXXX	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
XXXXXX	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0
XXXXXX	1	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
XXXXXX	0	0	0	0	1	1	0	0	0	1	1	1	1	0	0	0	0	0
XXXXXX	0	1	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0	0
XXXXXX	0	1	1	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0
XXXXXX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 3: *Screenshot of the primary data matrix produced by the information collected from the questionnaire.*



↓Contributor / Colu	Name	Surename	Gender	Age	Organization	email	Country	Use case title	Use case descrp
XXXXXX	XXX	XXXXX	Male	XX	NIH / NLM	xxxx@xxxx	USA	taxonomic	Connect taxonomic n
XXXXXX	XXX	XXXXX	Female	XX	UAegean	xxxx@xxxx	Greece	Migratory	Calculation of migratc
XXXXXX	XXX	XXXXX	Female	XX	NIMRD	xxxx@xxxx	Romania	Black Sea	Black Sea phytoplank
XXXXXX	XXX	XXXXX	Male	XX	USilKatowi	xxxx@xxxx	Poland	Checklist I	As part of the list is p
XXXXXX	XXX	XXXXX	Male	XX	MnhnL	xxxx@xxxx	Luxemburg	Taxonomic	Trying to managed ar
XXXXXX	XXX	XXXXX	Male	XX	UPorto	xxxx@xxxx	Portugal	InBIO Barc	DNA Barcoding datab
XXXXXX	XXX	XXXXX	Male	XX	DCLS	xxxx@xxxx	Japan	Duplicate c	We downloaded data

Fig. 4: *Screenshot of the metadata matrix produced by the information collected from the questionnaire*.

The univariate statistical analysis applied simple measures of the data which summarise the breath, the depth and the origin of the data and metadata collected, such as percentages of the data categories converted to pie-charts, bar-charts, etc. Although the results of the univariate statistical methods can directly be converted to knowledge statements, such as "*most of the users (e.g. 95%) require collection (specimen) data for their research*", they only make use of the data included in a single variable, they suppress all the information in that column in a single number (e.g. average value of the variable) and they're not capable of offering patterns and potential agents for the formation of these patterns.

Multivariate statistical methods can make use of all the data included in all variables taken into account during the research. They can also produce patterns in the form of plots, which can provide us with evidence on which variables correlate with others. In addition, multivariate statistics can provide evidence to infer which variables of the metadata (or other types of data) show patterns associated with those of the data. Both types of evidence can ultimately be converted to knowledge. The multivariate statistical analyses used during this study were: (a) the non-metric multidimensional scaling (nMDS) for the exploration of patterns of the data collected by the questionnaire (Kruskal & Wish 1978); (b) the BIOENV analysis for the association of the patterns derived from the data and metadata (Clarke & Ainsworth 1993).

To derive a similarity pattern from the above matrices, the Steinhaus coefficient was calculated between every pair of the variables of the above matrices. The values of this coefficient have formed the triangular similarity matrices which were subsequently used in the nMDS analysis in order to explore any multivariate similarity pattern. Finally, the resulting multivariate patterns were compared by using the BIO-ENV method. According to the mathematical procedure of the method, a harmonic rank correlation coefficient was computed between the two matrices (primary and metadata matrix). This coefficient may take values from -1 to 1. The value of the harmonic rank correlation coefficient shows the degree of association between the two matrices, that is, between their multivariate patterns. When the value is very high (0.9) then it can be inferred a high degree of association between the matrices. When the value is close to 0, then the degree of association inferred is very low. Finally, when the value of the coefficient is close to -1, then the matrices are inversely associated, indicating reversing patterns. The R Virtual Laboratory of LifeWatch ERIC (available by LifeWatchGreece; running in the HPC of IMBBC, HCMR) has been used for multivariate analysis.

3. Results

This section shows the results obtained from the questionnaire. It includes the profile of the researchers who participated in the survey, the origin and the type of data used by the researchers who replied to the questionnaire. It also includes their requirements and preferences. The relationships between their responses have also been analysed.

3.1. Profile of researchers interviewed

Data has been gathered from the profile of the researchers that responded to the questionnaire. It includes their gender, age and country of origin.

The majority of data comes from male researchers (52%). As it can be observed in Figure 5, female respondents represent only 19% of the total. As much as 29% of the respondents reported “other” in the gender choice or they did not specify their gender at all.

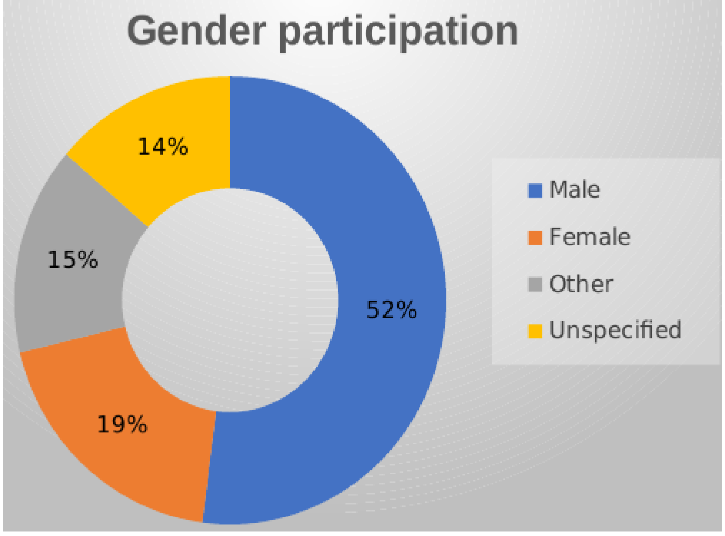


Fig. 5: *Gender percentages of the researchers that participated in the* BiCIKL *questionnaire.*

Regarding the seniority of the respondents, most of them belong to the most productive age class of 36-55 years old (58%). The participants are equally distributed between age class 36-45 (29%) and 46-55 years old (29%). This is closely followed by the class of 56-65 years old (23%).

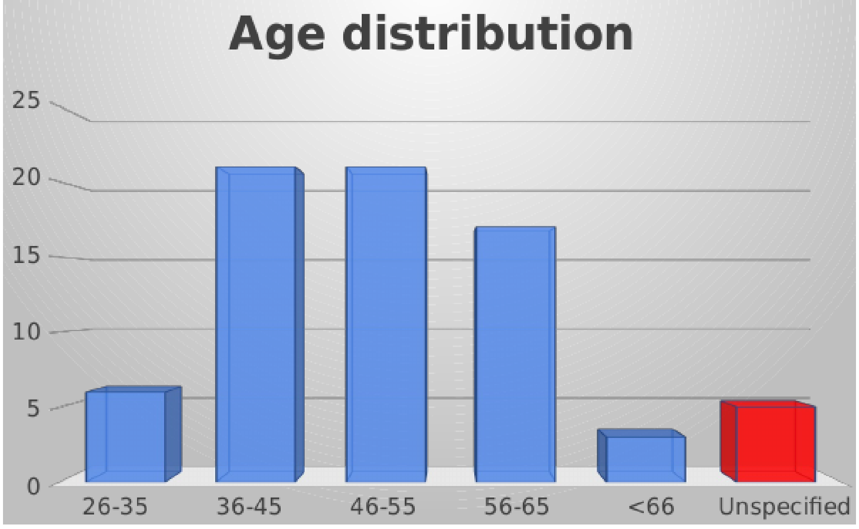


Fig. 6: *Distribution of the responding researchers to the* BiCIKL *questionnaire into age classes.*

The sample of the respondents include people from 30 countries, the top 5 of them being: UK (7 replies), France (6), USA, Bulgaria and Italy (5). The replies from these countries are closely followed by Belgium (4). With the exception of the USA, all of the above-mentioned countries are European ones. In total, only 33% of the researchers that provided responses come from outside Europe.

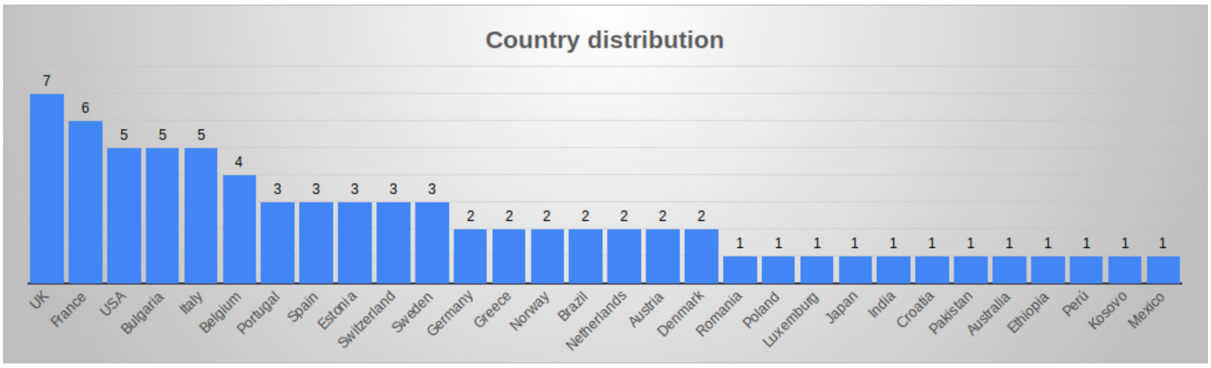


Fig. 7: Numbers of responses provided by the researchers aggregated by the country in which they work.

3.2. Types of data

Most of the researchers replied that they need *specimen* data (80%). This percentage is closely followed by the *taxon names* (75%) while *images* (63%) and *personal data* (56%) are also broadly required.

On the contrary, the type of data that is least required by the respondents are the *population parameters* (16%) and *big data* through a *cloud-computing* provider (19%). Half of the researchers require *molecular data*, *taxon treatments*, *sampling data* and *species traits* for their work. An important part of the respondents require *Operational Taxonomic Units* (34%).

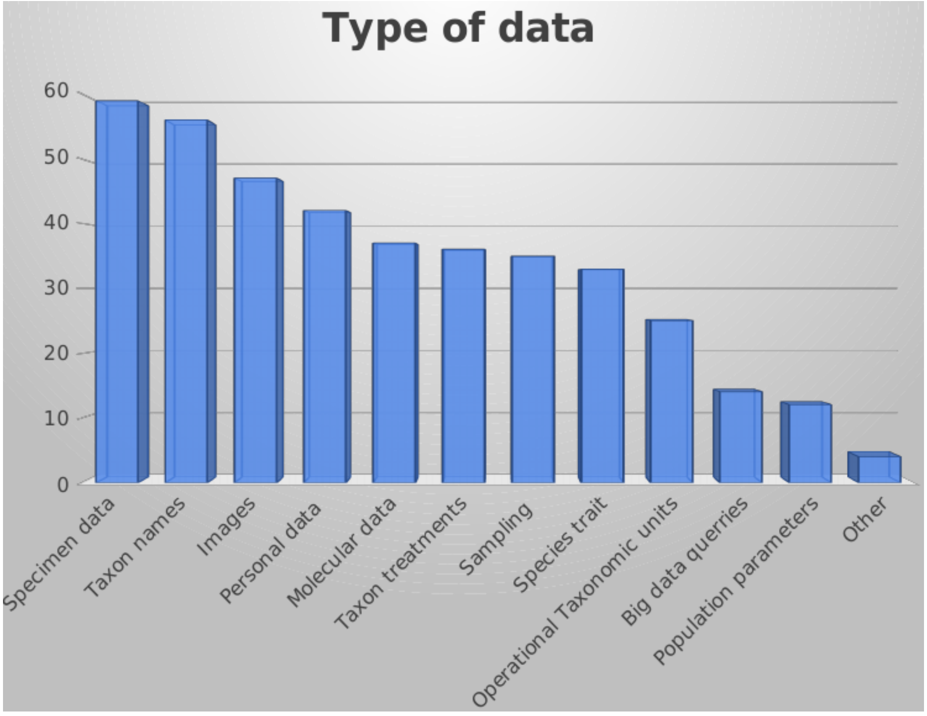


Fig. 8: Number of responses (y-axis) against type of data (x-axis) required from researchers that participated in BiCIKL questionnaire.

3.3. Data sources

The most required source of biodiversity data is the Global Biodiversity Information Facility (GBIF). GBIF is followed by BHL and Barcode of Life Data System (BOLD). On the contrary, the least required data sources are the Linked Open Data sources OpenBiodiv and SIB.

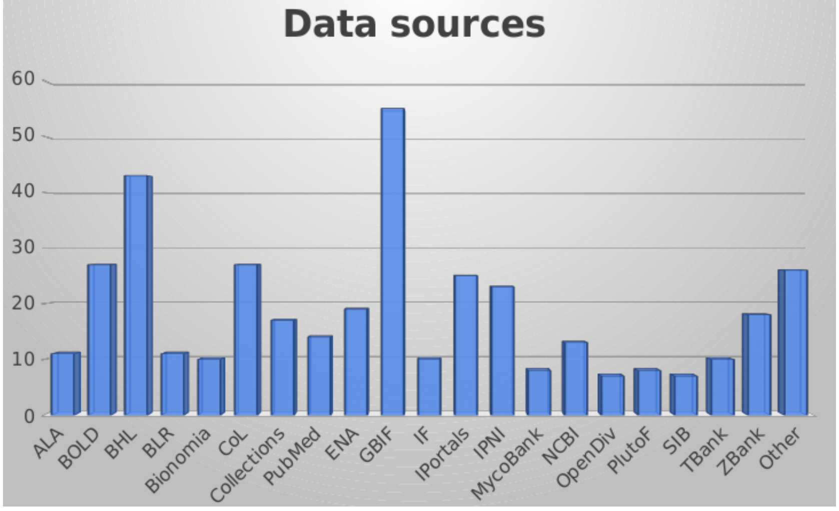


Fig. 9: *Number of responses (y-axis) against sources of data (x-axis) required from researchers that participated in BiCIKL questionnaire.*

3.4. Researchers preferences in using online tools

Most of the researchers who participated in the BiCIKL questionnaire declared that they have used or plan to use Wikidata (57%). Similarly, most of the respondents use their ORCID address for other purposes than publications (62%). The majority of the researchers declared they need standardised data (89%) and they make use of cloud services (94%).

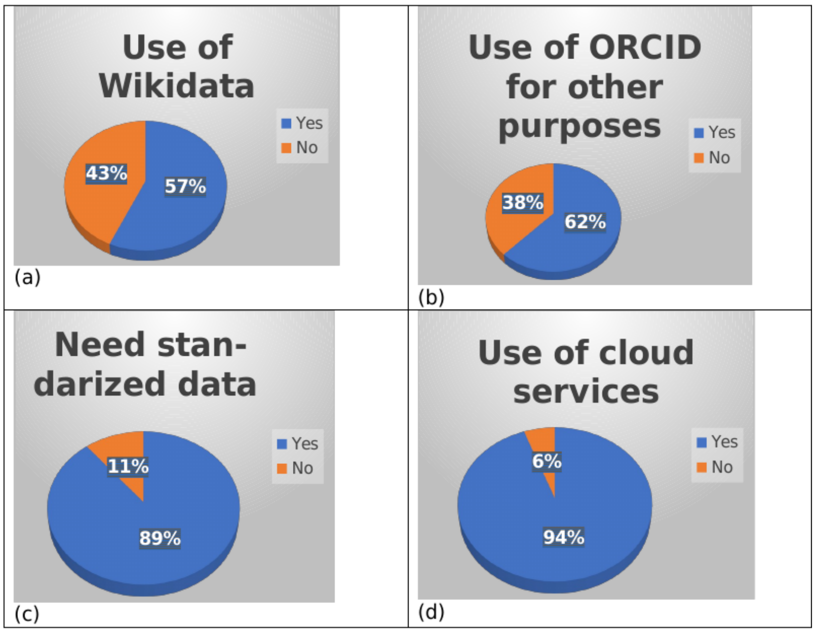


Fig. 10: *Use of Wikidata (a) by the researchers; ORCID ID for other purposes than publications (b); need for standardised data (c) ; cloud services (d).*

An important part of the researchers (38%) have reported problems with data infrastructures.

3.5. Multivariate patterns

The nMDS analysis, based on the matrix with the researchers who have replied to the questionnaire as rows and the question variables as columns, resulted in the following plot (Figure 11).

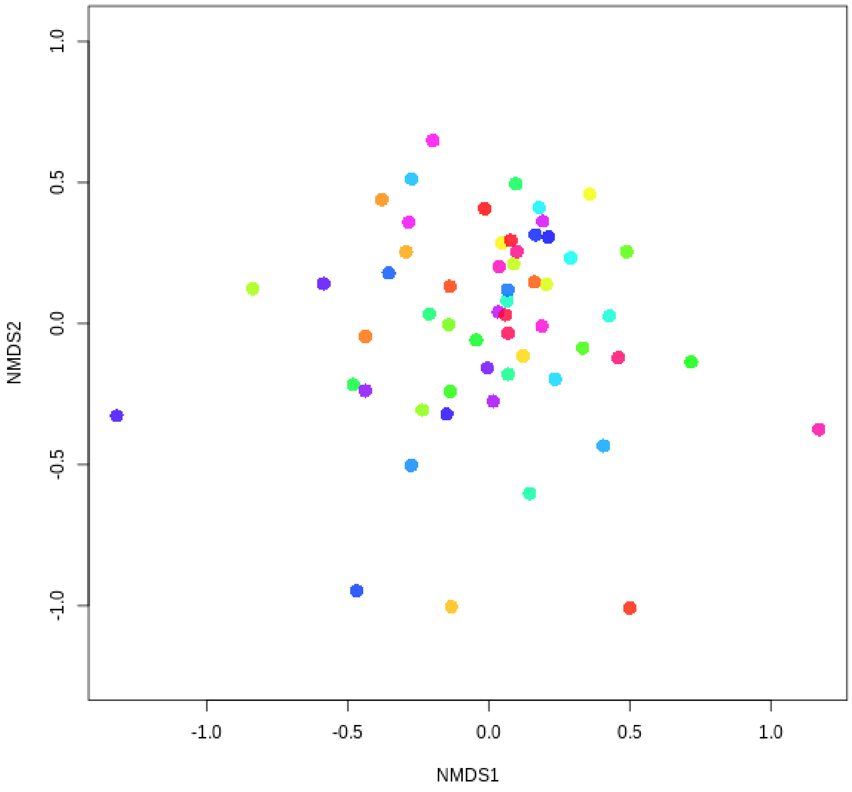


Fig. 11: *nMDS plot showing the multivariate pattern deriving from the similarities between the researchers' replies to the question variables of the questionnaire. Response IDs are represented by coloured dots.*

In the above plot, there isn't any significant pattern deriving from the replies the researchers gave to the question variables of the questionnaire. This means that we cannot differentiate any groups of researchers who share in common their requirements for types of data, quality and access modes to them. It is, therefore, interpreted that their requirements are individualistic and subjected to the kind of research they do (or intend to do).

When we transpose the matrix so that the rows become the question variables and the columns the researchers who have replied to the questionnaire (IP addresses), the analysis results in the same pattern, as above: no groups of question variables sharing in common the preferences of the researchers who replied to the questionnaire can be found in the plot (Figure 12).

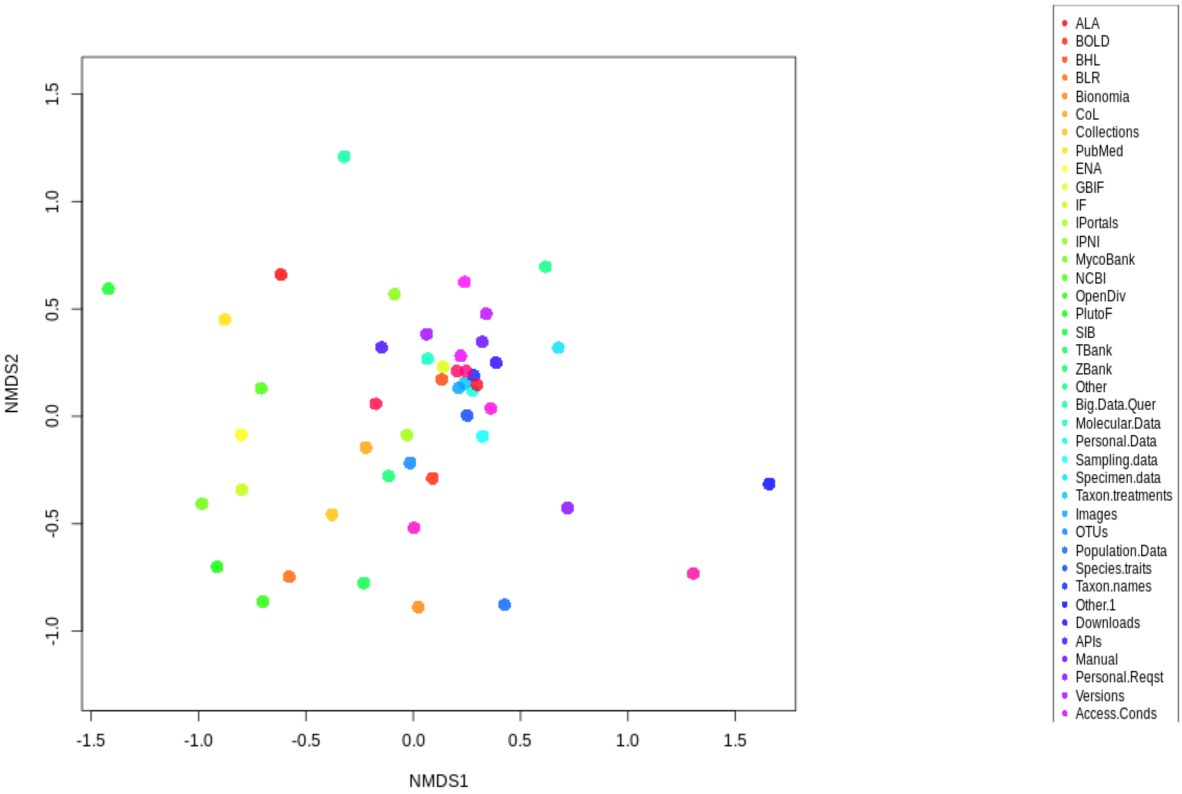
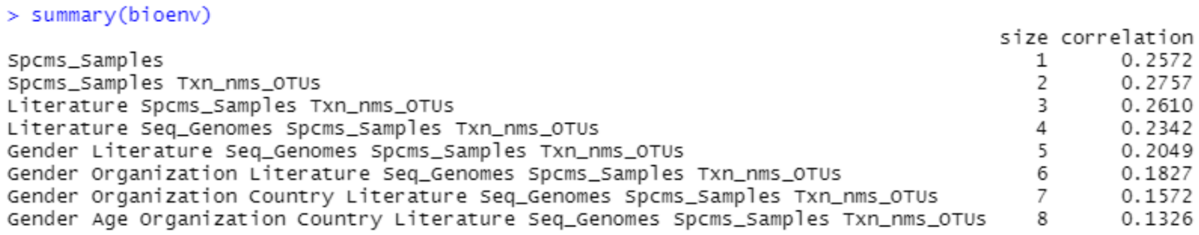


Fig. 12: *nMDS plot showing the similarities between the question variables as deriving from the researchers' replies to the of the questionnaire*.

The BIOENV analysis resulted in very low correlation values (max. 0.2572), which was an expected result since there was no particular pattern in the requirements of the researchers. Therefore, there are no specific variables which may be associated with the researchers requirements (Figure 13).



```
> summary(bioenv)
                                                                              size correlation
Spcms_Samples                                                                    1      0.2572
Spcms_Samples Txn_nms_OTUs                                                       2      0.2757
Literature Spcms_Samples Txn_nms_OTUs                                            3      0.2610
Literature Seq_Genomes Spcms_Samples Txn_nms_OTUs                                4      0.2342
Gender Literature Seq_Genomes Spcms_Samples Txn_nms_OTUs                         5      0.2049
Gender Organization Literature Seq_Genomes Spcms_Samples Txn_nms_OTUs            6      0.1827
Gender Organization Country Literature Seq_Genomes Spcms_Samples Txn_nms_OTUs    7      0.1572
Gender Age Organization Country Literature Seq_Genomes Spcms_Samples Txn_nms_OTUs 8     0.1326
```

Fig. 13: *Results of the BIOENV analysis with the highest correlation coefficient values, indicative of a very low association between the multivariate pattern of the researchers' requirements and that of the relevant metadata*.

These results of both the nMDS and BIOENV analysis are indicative of an unbiased sample, that is, in the researchers' requirements who participated in this study.

4. Conclusions

The univariate statistics shows that the majority of the participating researchers: (a) are males; (b) are in the most productive age classes; (c) work in European countries; (d) require *specimen data*, *taxon names* and *images* for their work; (e) use mostly GBIF, BHL and BOLD as data sources; (f) use Wikidata, cloud services and use their ORCID ID for purposes other than publications; (g) need standardised data.

However, these preferences are randomly distributed in the sample of the researchers who participated in this study. The multivariate statistical analysis has confirmed the above randomness in the researchers' preferences by demonstrating no specific pattern in the requirements of the researchers in the types of data, quality and access modes. In addition, there seems to be no specific metadata associated with the above pattern. These results also indicate that there are no biases in the sample of the researchers who participated in this study.

5. Acknowledgements

The authors would like to acknowledge the BiCIKL Consortium for their support of the activities carried out in the course of this study. They also feel much indebted to all the researchers who replied to the questionnaire and provided the BiCIKL Consortium with such valuable information and data. The EC is specially thanked for the financial support to this research: the BiCIKL project is funded by the EU (European Union’s Horizon 2020 Research and Innovation Action under grant agreement No 101007492). Finally, special thanks go to Dr. Christina Pavloudi (University of George Washington, Washington DC, USA) and Mr. Savvas Paragamian (IMBBC, HCMR, Heraklion, Crete) for running earlier versions of data with the multivariate statistics.

6. References

1. Clarke KR, Ainsworth M (1993) A method for linking multivariate community structure to environmental variables. *Marine Ecology Progress Series* 92:205–209
2. Kruskal, J.B., Wish, M. (1978) Multidimensional scaling. Sage Publishers, Beverly Hills