ELGO-DIMITRA Data Management Practices & Requirements:

A Scoping Report

Table of Contents

ELGO-DIMITRA Data Management Practices & Requirements

Version 1.0, October 2024

Authorship and citation information

The report was jointly authored by Fotis Tsiroukis and Sabina Leonelli, in collaboration with ELGO-DIMITRA. Please cite as follows:

Tsiroukis F., Leonelli S. and ELGO-DIMITRA (2024) *ELGO-DIMITRA Data Management Practices & Requirements: A Scoping Report*. PHIL_OS Report. DOI: 10.5281/zenodo.13999260

Dataset DOI: 10.5281/zenodo.14003418

Acknowledgments

This work has been supported by the PHIL_OS project ('A Philosophy of Open Science in Diverse Research Environments', www.opensciencestudies.eu), funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001145). We express our gratitude to ELGO-DIMITRA staff for invaluable support and contributions, as well as PHIL_OS team members at the University of Exeter for precious feedback on early drafts.

Foreword

As the commissioner of this study on data practices within the Hellenic Agricultural Organization DIMITRA (ELGO-DIMITRA), I am pleased to have in hands this original work of reporting the data management practices within our organization. Despite the key role of ELGO-DIMITRA in applied agricultural research, this is the first time that such a topic is studied in close cooperation with Prof. Leonelli's research group, PHIL_OS. The present report is a crucial step for recording the current status and highlighting the existing problems, a prerequisite for a more effective planning of improvements, in both hardware and strategy, within ELGO-DIMITRA. Given the importance of storage and re-use of existing data for purposes beyond the original scope of their collection, we believe that the long-term benefits of this effort are crucial for achieving a more efficient transition from sets of strictly scientific data, to scientifically-proven decision and policy-making tools, leading to improvements from farm-scale to regional or national level.

Dr Anastasia Tsagkarakou
General Director of Agricultural Research, ELGO-DIMITRA

1. Introduction

This study provides insights into the current state of data practices in agricultural science as carried out within the Greek Agricultural Organization ELGO-DIMITRA. ELGO-DIMITRA is a central national governmental organization with three main roles: a) to conduct research in agriculture locally, b) to guarantee best quality of agricultural products through certifications and standardization of genetic plant resources and cultivation practices and c) to provide education to farmers and producers. This report aims to foster appropriate investment in long-term data infrastructures and data-related skills as well as a deeper awareness of the richness of Greek research practices and related data collections within the international landscape.

Good science needs reliable and effective practices of data management, including data collection, storage, processing, circulation and re-use. This is especially important for the fields of agricultural research, crop science and plant science which deal with local varieties, in a context that is bound by the sociopolitical, economic and material conditions of the nation state under which research operates, as well as its unique geographic, ecological and climatic conditions. The amount and variety of data relevant to agricultural research is high, involving information about plant physiology, microbiology, genomics, geophysics, hydrology, ecology, entomology, biochemistry, climate science, socio-economic indicators and food quality, among other relevant domains. Linking and integrating these data towards scientific investigation is a challenge, and adequate data management and stewardship are integral to tackling it.

Such is the case of agricultural research in Greece, a region that is affected by climate change and economic precariousness. Research data in this context play a pivotal role well beyond academic research: the data can inform policy-making, local farming practices, land management, urban planning and business decisions. For research to have a meaningful impact in the local social context as well as in tackling global climate change effects, practices of data collection, storage, analysis and sharing need to be supported by strong foundations in research management, responsible practice, digital and material infrastructures and relevant education.

ELGO-DIMITRA is a pivotal organisation for Greek agricultural research. It comprises of 11 institutes, with a central coordinating body in Athens. These institutes are distributed all over Greece and are mainly centred around Thessaloniki, Athens, East Thessaly and Crete. Each institute is composed of other departments with a broader geographical reach, both in continental areas and Greek islands. The disciplinary range of the institutes is wide-ranging, integrating research in local crops, water management, soil, plant genetics, forestry, animal husbandry, fisheries, ecology, economics and sociology among others. This study concerns data practices within the 11 institutes, as self-reported by its scientific staff by means of a survey conducted in late 2023, and thereby documents some of the key opportunities, challenges and potential for Greek research of agricultural relevance going forward.

2. Methods and Sample Analysis

Methods

This study took the form of a survey distributed amongst researchers and research-adjacent employees of HAO amongst its various institutes around Greece. The survey consisted in a text-based online form listing 20 questions (including consent in the beginning and an optional open comment section in then end). The survey was designed by members of the PHIL_OS research project in partnership with the ELGO-DIMITRA General Directorate of Agricultural Research. The Directorate took on the distribution of the survey form within ELGO-DIMITRA research institutes. The survey became available online on November 20, 2023 and was distributed to potential respondents in the following week. It was open to responses over the course of the following month, receiving its last response on December 27, 2023 (4:30 PM EET).

Dissemination

The survey was administered through an anonymous link, preserving the anonymity of the original responders by not tracing emails or IP addresses. The survey link and instructions for participating were disseminated by the general director to all the researchers through email, and encouraged them to disseminate it to the members of their groups. Two reminders have been sent subsequently to researchers and institute directors in charge of cascading the survey down to their staff locally. Given the level of anonymity within the survey and the fact that participation in the study was voluntary, it is not possible to determine in which degree staff from each institute responded.

Design

The survey's questions, their content and their phrasing were developed in close collaboration with the General Director of Agricultural Research, as well as the Director of the Institute of Olive Tree, Subtropical Crops and Viticulture in Chania. Therefore, the questions reflect the interests and motivations of the ELGO-DIMITRA Directorate of Agricultural Research and are close to the experience of working in ELGO-DIMITRA as a researcher. Moreover, the general director was involved in providing feedback related to survey flow and design choices after the questions were finalized and entered into the web-based survey. The final decisions about feedback and edits from the Directors were taken by the main authors of this report (Fotis Tsiroukis and Sabina Leonelli), who reviewed them, edited them and created the web-based survey.

The survey was designed on the web-based platform Qualtrics and included a variety of question types such as multiple choice, ranked choice, scales, yes/no and free text. The variety in the types of questions allows for both a quantitative and qualitative analysis of the survey data and provides a richer and more complete picture

of data management practices and needs on the ground. For example, 3-point scale questions such as Question 4 (Q4) allow for the identification of statistical patterns related to the choice to share data amongst a diversity of data types from various disciplines and in various formats, while the ability to give free text answers to the choice "Other" allowed for the integration of a greater variety of data types that were not initially anticipated. The mixed methods approach of the survey also allows to make cross-comparisons between quantitative and qualitative data to provide more depth to the qualitative results and to be able to account for interesting patterns identified in the data. This helps to provide better informed suggestions and crowdsourcing suggestions through the free text answers.

Quantitative Analysis

Qualtrics automatically performs basic statistical analysis in terms of percentages, means and medians which provides the backend data for constructing visual representations. This means that for the multiple-choice, yes/no, scale and ranked choice questions, Qualtrics handled the quantitative analysis. For the free text questions, a mixture of quantitative and qualitative analysis was performed. For example, for some questions which included lots of repeated names and words, keyword frequency analysis was performed on the raw text data. A sheet of the raw data from all responses was made and additional pages were made on Microsoft Excel for clearer formatting of the cells and for conducting additional analysis if needed. For questions Q6-Q9 the formatted data was fed to ChatGPT for keyword frequency analysis and calculating percentages.

The variety of question formats also allows for a range of cross-comparative additional operations to be performed outside of the per-question analysis present in the Appendix. In fact, for some interpreting some of the results of some questions, such an analysis was needed. There remains scope for further and more detailed statistical analysis of the dataset in the future.

Qualitative Analysis

Thematic analysis was performed to free text questions in conjuction with some rudimentary qualitative analysis. Open coding, thematic categorization and clustering into main themes was performed with the aid of ChatGPT and prompt engineering to fine-tune answers and account for errors and mistakes from the generated answers. The validity of ChatGPT responses was cross-checked with the raw data. Themes generated from thematic analysis with GPT have also undergone significant modification and editing.

Sample Analysis

The survey was designed for and distributed to the **182 research staff** working in the eleven institutes belonging to ELGO-DIMITRA all around Greece, as follows:

Northern Greece (Thermi, Kavala)

1. *Institute of Fisheries Research*
2. *Institute of Forest Research*
3. *Institute of Animal Husbandry Science*
4. *Institute of Veterinary Research*
5. *Institute of Plant Breeding and Genetic Resources*
6. *Institute of Soil and Water Resources*

Central Greece (Larissa, Volos, Patras, Karditsa)

7. *Institute of Industrial and Forage Crops*

Athens

8. *Institute of Mediterranean Forest Ecosystems*
9. *Institute of Technology of Agricultural Products*
10. *Institute of Agricultural Economics Research*

Crete (Chania, Heraklion)

11. *Institute of Olive Tree, Subtropical Crops and Viticulture*

Some of these institutes have departments located around other areas of Greece. For example, the *Institute of Olive Tree, Subtropical Crops and Viticulture* (IOSV) which is based in Chania (Crete) also has departments in Heraklion, Athens, Kalamata and Lesvos covering a big array of different regions of Greece with diverse climatic and cultural conditions.

The survey was distributed over email by the director of ELGO-DIMITRA and respondents were given the option to fill it in on a voluntary basis. From this initial sample of 182 potential respondents, the survey had a total of 103 responses, **70** of which had a 100% completion rate. Because of the anonymous nature of the survey's distribution and its top-down dissemination, it is not clear which institutes are represented and how many respondents were from each institute.

Demographics

Looking at the answers of the participants we can get a more accurate picture of the sample and what demographics it represents. The most important of these is depicted by the following graph which represents the highest qualifications of the sample (from *Q18.*):

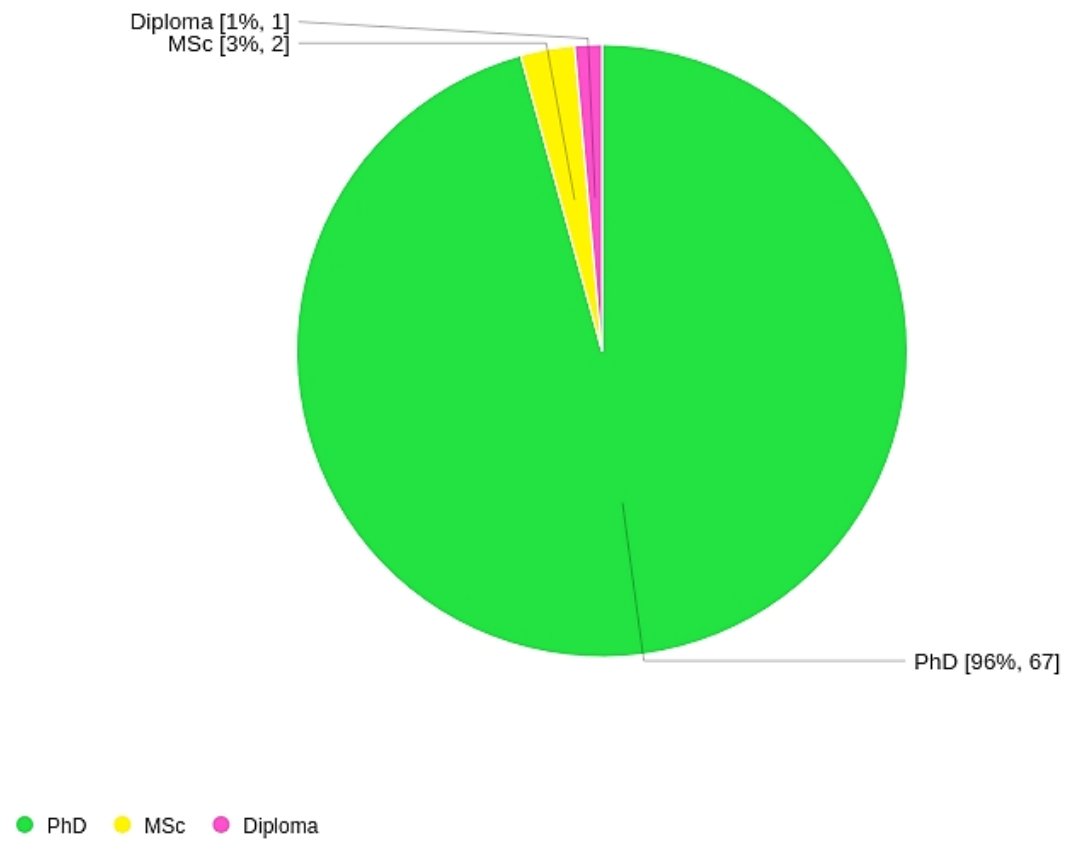


Figure 1: Q18. What is your highest qualification?"

The sample is composed mostly of PhD holders with only a tiny fraction (4%) holding MSc degrees or other. This indicates that the level of education, specialization and expertise in the sample is high, which is not surprising given that the potential respondents are mainly professional researchers working for ELGO-DIMITRA and to a lesser degree the supporting personnel of their research groups.

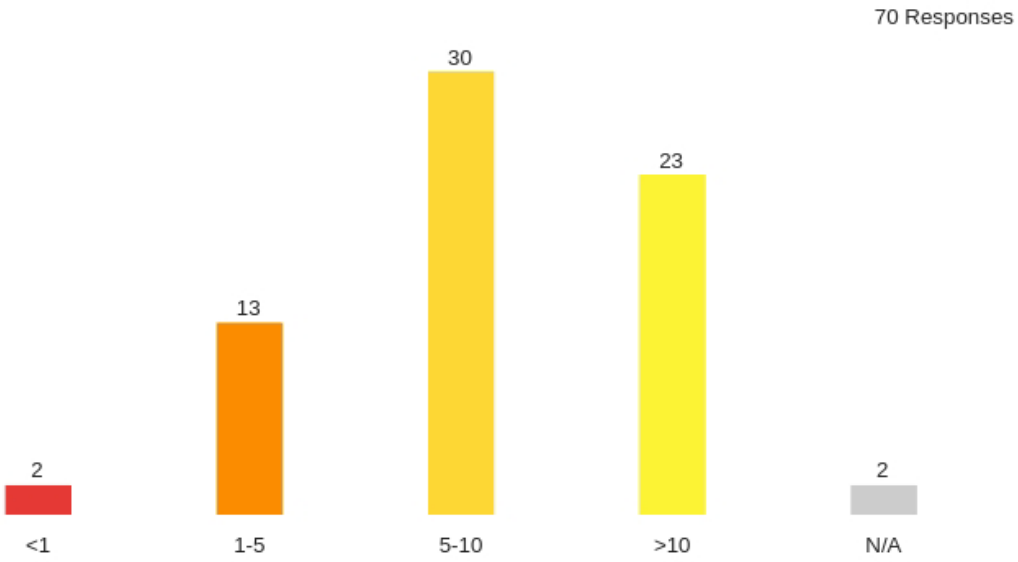


Figure 2: Q16. "How many years have you been working at ELGO-DIMITRA?"

Moreover, the results (from Q16.) show that most of the respondents have been in the institute for >5 years and 23% has been for more than >10. We can directly infer from these results that many of the respondents are not early career researchers and that the institutes are composed mainly of more experienced researchers.

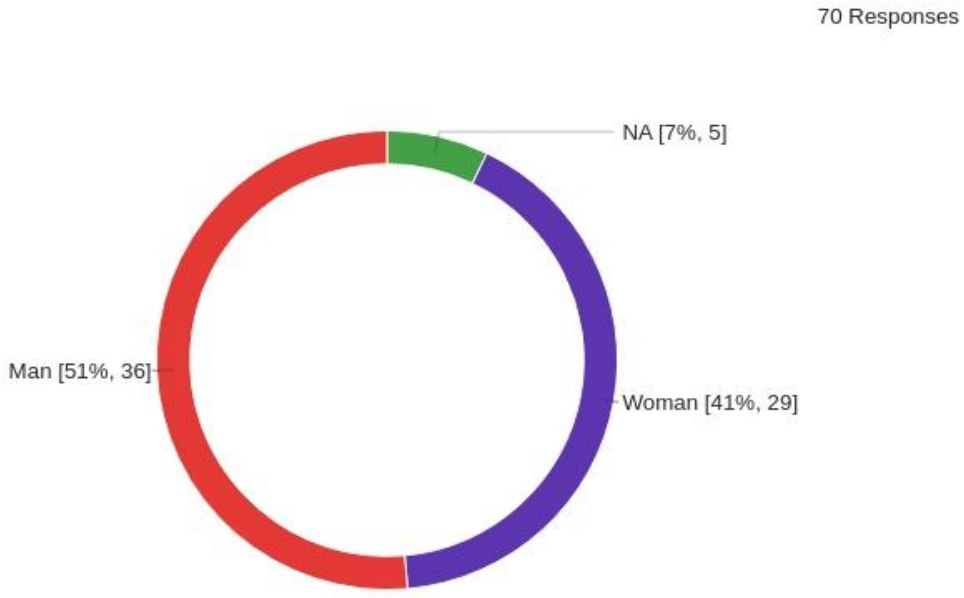


Figure 3: Q17. "What is your gender?"

Regarding gender distribution, the results show the most (51%) respondents are men but the percentage of women is not lagging too far behind (41%). The distribution is not

equal but the difference is not indicative of a big disparity. However, women are mostly found in lower ranking roles. When it comes to the other 7%, N/A, refers to an option “Prefer not to say”, while there was also option for “Other” which nobody chose.

An additional component of the demographics that is specific to the context of ELGO-DIMITRA and Greece is researcher ranking. The Greek Ministry of Agriculture sets a well-defined hierarchy of ranks with different expectations, responsibilities, evaluation criteria and privileges. It is composed by three formal ranks from A-C (higher-lower) as well as the rank of the Director which has a more administrative status. A few researchers in the organisation were still in rank D level during the survey, although this refers to the previous 4-level ranking of the organization at the time they were hired, which is non-existing under the current legislation. The rest are contractors and support staff:

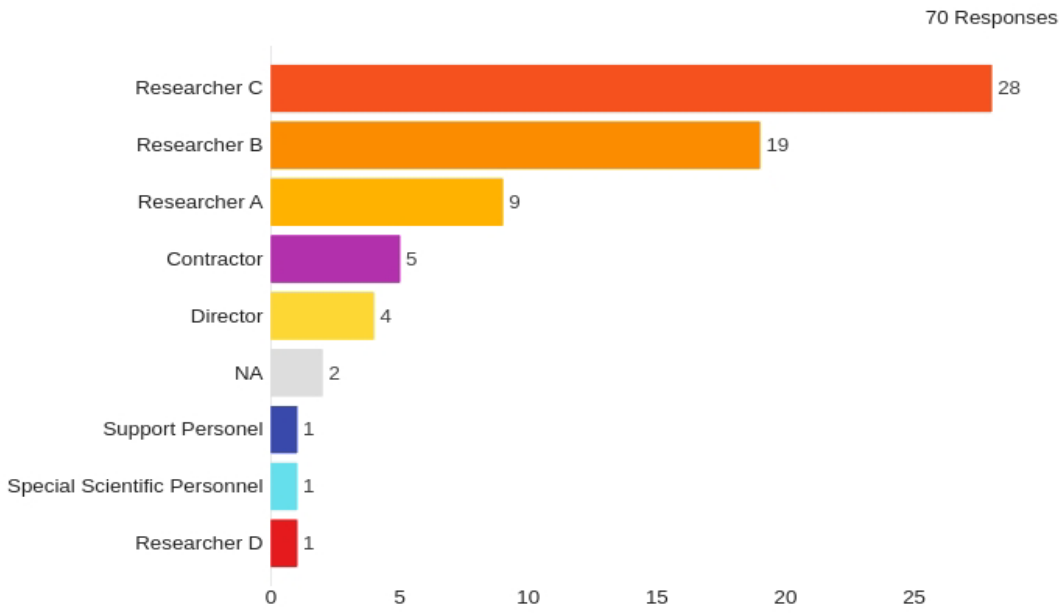


Figure 4: Q15. "What is your role in ELGO-DIMITRA?"

Most respondents pertain to ranks C and B, which comprises the majority in the sample. Contractors and support stuff represent only a small fraction. The number of directors in the respondents is not a reliable indicator of the number of institutes who responded, since some directors might not have completed the survey because of time pressures.

It might be the case that younger researchers, at assistant roles, working as contractors did not feel as motivated to complete the survey or the survey did not get distributed to these positions. More established researchers may have more systematic knowledge of the data practices at labs and thus might feel more confident to answer.

According to the legislation, the minimum requirements for ranking researchers in each level are as follows:

- ➔ *For rank C*: The researcher must have documented ability to design and execute research and technological development projects or parts of projects, and to allocate parts or phases of the project to other researchers, providing guidance or supervision. Additionally, he/she must have original publications in international peer-reviewed scientific journals.
- ➔ *For rank B*: the researcher is required to have documented ability to organize and manage research and technological development programs, coordinate and direct research and technological development in individual projects under the frame of the Research and Technological Development program, seek and receive funding from external sources for the activities of the institute or the organization, and promote pioneering ideas in science and technology. It is also required that the researcher has made original publications in international peer-reviewed scientific and technical journals or has international patents and has contributed to the advancement of science and technology, with their contributions recognized by other researchers.
- ➔ *For rank A*: the researcher must have a proven ability to develop research and its applications in new fields, coordinate activities in broader areas of the Research and Technological Development (R&TD) program, contribute to the formulation of research and technological policy, and the development of research organizations by receiving external funding. They should be internationally recognized for their contribution to the advancement of R&TD in scientific and technological areas of their expertise, as well as for the dissemination and application of knowledge produced from research and they should have a significant number of publications, including monographs or original publications in international peer-reviewed journals and hold significant patents.

Limitations

Despite the good response rate and amount of detail in the free text data, there are certain limitation of this study that need to be addressed. These have to do with its dissemination, its design and the analytical methods used.

1. Dissemination and Scope

Responses should not be taken as fully representative of what respondents do in practice. A great deal of information that might be informative for the survey might not be reported for reasons of brevity and lack of time.

The survey included questions about the types of data shared but not about how many data and how often they are shared by researchers in different fields. It was therefore not possible to weigh responses against the volume of contributions and experience by each respondent.

The results about career stage and role (Q15.) show that technicians and supportive personnel were under-represented in the survey sample (4%). These roles may have different ideas of data management which might not be reflected in the sample.

Lastly, the sample of ELGO-DIMITRA employees should not be taken as indicative of the data management practices of agricultural science and plant science in Greece in general. Agricultural research is also conducted in university settings or by private for profit or non-governmental organizations (to a lesser extend). These provide a different context and conditions, which are not covered by this survey.

2. Survey Design

Respondents were not asked which institute they worked with, (to avoid the danger of de-anonymisation). This makes it hard to know if there are possible connections between the types of data shared and the research focus of a certain institute. Such information could have also been informative about whether certain problems are common or if they have to do with specific institutes. Such information would have made questions like Q4 & Q5 easier to interpret and properly contextualize. It would also provide a window into the diversity of the research environments around the institutes.

Question Q5 might have been difficult for respondents to parse, since it required ranking on a scale of 9 (corresponding to the number of data type choices). It was also challenging to visualise and interpret the resulting data, with horizontal bar charts proving the most approachable method but still containing limitations when it comes to colour-coding ranked data, especially when the data are of the same type.

3. Analytical Methods

This survey was designed in Qualtrics with related limits in capabilities for reports and data analysis. For instance, Qualtrics doesn't provide the ability to cross-compare data between questions and make graphs out of them. Thus, these operations have to be done outside of the software which makes the process time-consuming and more prone to error. Therefore, more advanced methods of analysis have been left out of the present analysis. The raw data are available in appendix for further analysis.

Excel-based analysis and ChatGPT have been used as complements to the limitations of Qualtrics when it comes to: a) free text data and thematic analysis, b) cross-comparisons of data from different questions. However, they come with their own limitations.

Using Excel requires exporting data from Qualtrics, but exporting raw data from Qualtrics and the way excel formats data cells can be a limitation for performing quantitative analysis on multiple choice questions where respondents choose multiple answers. The individual respondents' multiple answers are comma-separated which is challenging to parse with Excel. However a workaround has been found at least for Q2 (see Appendix Q2).

ChatGPT was used for automatic open coding, thematic analysis and keyword analysis, and the reliability of this tool depends on careful prompts and human checks. ChatGPT-generated codes and themes have been re-edited and curated afterwords by the report authors, but there might still be issues with patterns found in the data through the LLM algorithm.

Summary Table

<table>
<tr><th colspan="2">Methods</th></tr>
<tr><td>Medium</td><td>Online survey (text form) with a variety of question types:
multiple choice
ranked choice
scale
yes/no
free text</td></tr>
<tr><td>Distribution</td><td>Anonymous Link & Email in Mailing List (From Directorate to Institutes to Researchers)</td></tr>
<tr><td>Software</td><td>Qualtrics (Survey Design, Database & Visualization)
Excel (Formatting, Coding, Open Data)
ChatGPT (Free Text Quantiative Analysis, Thematic Analysis)</td></tr>
<tr><td>Analysis</td><td>Quantitative (Percentages, Keyword Analysis, Mean)
Qualitative (Thematic Analysis – Open Coding)
Mixed (Theme Frequency Analysis)</td></tr>
<tr><th colspan="2">Sample</th></tr>
<tr><td>Location</td><td>Greece (Athens, Northern, Central, Peloponnese, Crete, Aegean)</td></tr>
<tr><td>Sample Size</td><td>Target Sample: 182
Survey Sample: (103 partial responses, 70 fully completed)

The survey results are based on how many respondents have answered each question which can vary, depending on the percentage of completion of the survey.

Demographic Information

ELGO-DIMITRA Employees:
Researchers
Directors
Lab Technicians
Supportive Personnel
Gender:
51% male, 41% female

Qualifications:
96% PhD, 3% MSc, 1% Diploma</td></tr>
</table>

3. Main Results

For a detailed analysis of each individual question and their graphs, ***See Appendix****. This section will contain lump summaries and a comparative analysis of questions.*

Data Collection Practices

- Researchers from ELGO-DIMITRA employ a mix of traditional and digitized methods for data collection. While there is a prevalence of physical media such as paper and notebooks, a substantial portion of researchers opt for digital mediums, including Excel sheets, mobile devices, and various digital data management tools and software.

Data Storage Practices

- There's a discernible tendency for distributed solutions that necessitate an internet connection, indicating that there exists high reliance on cloud-based computing tools rather than local storage on local servers or institute devices.
- The choice of data storage platforms predominantly leans towards proprietary cloud services offered by American providers, with Google Drive emerging as the dominant preference followed closely by Dropbox. While there are few publicly funded alternatives like Zenodo in use, they are not as systematically adopted.
- Most selected services operate on a freemium model, where access to basic features is free, but for access to more nuanced features (that the researchers might need) a paid plan is needed. Additionally, a minority of researchers opt for citation managers such as Mendeley or open web platforms like Zenodo for data storage.

Data Sharing Patterns

- Data sharing among researchers within the ELGO-DIMITRA organization exhibits distinct patterns. While a vast majority of respondents share their data with collaborators outside their lab, intra-departmental sharing within ELGO-DIMITRA is neither systematic nor frequent.
- Notably, almost all researchers engage in some form of data sharing, with approximately 62% choosing to share with other national and international institutes. Only a small fraction, constituting 3%, refrain from sharing beyond their own lab.
- Moreover, of those who do share externally, a mere 6% opt to share with other ELGO-DIMITRA departments, indicating that they prefer to complete data analysis and publish their work first.

Data Types and Sharing Decisions

- Sharing preferences regarding highly specialized data types underscore the nuanced nature of data dissemination practices among ELGO-DIMITRA. While a significant portion, amounting to 63%, opt for sharing under specific

circumstances, a notable proportion falls under the "Other" category, indicative of the diverse array of data types encountered across various disciplines within the institute. Interestingly, researchers demonstrate a heightened propensity to share highly specific data types, such as biometrics of forest trees or bee population measurements, despite the laborious nature of obtaining, processing, and curating such data.

- For a more detailed account of the exact specialized data types that respondents gave, see Appendix Q4.

Data Management Practices, Challenges and Support

In terms of data management practices, researchers employ a multitude of storage platforms, encompassing both cloud-based services and local storage solutions.

- The primary challenges faced revolve around infrastructural limitations, including issues related to volume, speed, and the lack of robust backup mechanisms. Notably, some respondents attribute these challenges to specific data types, indicating the need for tailored solutions to address domain-specific requirements effectively.

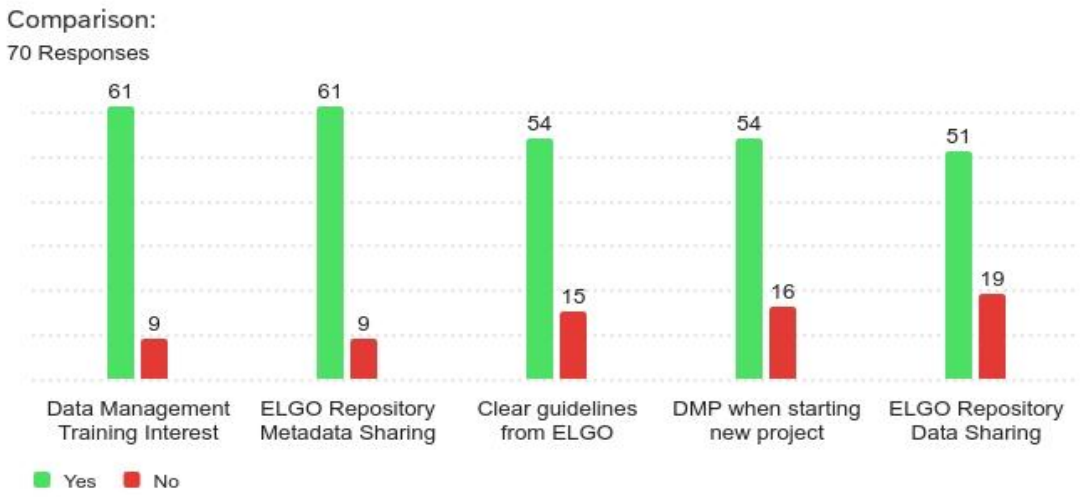


*Figure 5: Comparison between questions on Data Management **Q10-Q14***

- ELGO-DIMITRA researchers exhibit a willingness to engage with institutional support mechanisms for data management. This encompasses training, guidance, and the provision of data-sharing infrastructures by the organization. Moreover, there's notable interest among researchers in improving data management practices, as evidenced by the high percentage of respondents expressing interest in data management training and the preparation of Data Management Plans (DMPs) at the outset of their projects. This underscores the growing recognition of the importance of structured data management practices within the research community.

Conditions for Data Sharing

- Researchers demonstrate varying degrees of openness to sharing their data, particularly concerning repositories managed by HAO. While there is a general willingness to share data openly, the conditions for sharing vary, with some researchers preferring a permissioned system of access, wherein prospective users are required to submit information about their intended use of the data. This approach reflects a balance between fostering open access to data and ensuring responsible data stewardship practices within the organization.

Summary Table

Data Collection	**Mixed analog/digital means of data collection:** - Paper and Notebooks (30%) - Digital Media (Excel Sheets, Mobile Devices, Software) (70%)
Data Storage	**Distributed/cloud storage over Local Storage (servers)** - Cloud Services (Google Drive, Dropbox) (80%) - Proprietary vs. Publicly Funded Alternatives (e.g., Zenodo) - Freemium Models vs. Other Choices (e.g., Mendeley, Zenodo)
Data Sharing Patterns	**Sharing Patterns:** • Overall number of sharing outside one's lab: 97% • Sharing with national/international institutes: 62% • Sharing within ELGO-DIMITRA departments: 6% • No sharing: 3% **Decision-Making Factors:** - Degree of Specificity of the Data Types - Degree of Completion of the Project - Control over Conditions for Sharing
Data Types and Sharing Decisions	**Variety of specialized data types** - Sharing under specific circumstances: 63%
Data Management Practices, Challenges and Support	**Challenges:** - Infrastructural limitation: volume, speed, lack of robust backup mechanisms. - Specific issues related to handling specialized data types **Openness to Support:** - Interest in Data Management Training: 87%
Infrastructural Conditions for Data Sharing	More willingness to share under a system of permissioned access: 86%

4. Discussion

The survey evidences four key trends in the data management practices of the institute, as follows:

1. ***Data sharing under agreed conditions* is preferred to unconditional data sharing**: Researchers display a preference for well-defined, restricted conditions over the opportunity to share data unconditionally. This trend is particularly strong (three times as likely to share under restricted conditions) in the case of highly specific types of data which may be laborious to obtain (e.g. biometrics of forest tress, bee population measurements). As indicated by comparison between Q12 and Q13, researchers have a strong preference for using data repositories to upload information about their data (metadata) rather than the data themselves. This strategy makes data findable and requires anybody interested in accessing the data to contact the original data producers, thereby generating fruitful dialogues between research groups. The strategy fits a model of "Open Data" that facilitates judicious connections (Leonelli 2023), while at the same time complying with the FAIR data principles by making data Findable, Accessible, Interoperable and Reusable.

2. ***Data storage and collection are multifaceted and diversified*:** When considering the ensemble of responses concerning preferences around data collection and storage, it becomes clear that respondents use a variety of tools and strategies in ways that are complementary and multifaceted. Analog/physical and digital options are often chosen together. This may indicate the significance attached by researchers to ensuring that all possible forms of storage are covered, to prevent data loss. It may also indicate a moment of transition towards the digitalization of research practices and outputs, with data being collected on paper and only later entered on a spreadsheet or other digital tools and uploaded online. More research, including qualitative methods such as interviews, is needed to determine the reasons for the current diversity of modes and media in data handling.

3. **Data sharing happens through local as well as distributed tools:** There are two different approaches to data sharing that respondents use the most. The first consists in storing data *locally fist* by keeping digital copies and modifications of these copies that are stored on the person's device and synced when they have internet access. This encompasses the use of personal devices, institutional computers, local server databases etc. The second consists in storing data in distributed ways on a server first, without the need to store locally. This requires an internet connection and for the user to be logged in to an account. It also requires access through a browser or desktop app with reliable internet connection. Examples of these are cloud services or remote repositories. The results of Q6 show a definite preference of distributed cloud storage technologies while a comparison with the results of Q3.1 show that proprietary storage technologies are the ones being used the most. **The data**

shows that most researchers desire an institutional remote repository with server infrastructures in place, which would be an intermediate choice providing both the benefits and flexibility of remote access while at the same time providing the data sovereignty and resilience benefits of a local institutional databank.

4. **Several different technologies may be considered appropriate for digitalization**:

   The majority of responses to Q1 included >3 choices and amongst them a significant number of respondents include both paper and digital media. This indicates that researchers are not choosing to use paper simply because they lack other, digital means for data storage and sharing. Rather, the use of paper tools indicates methodological diversity and plurality in the labs, where different media and instruments are used. Understanding the needs, goals and motivations underpinning such diversity, and the reasons for choosing data collection media at each stage of research, requires follow-up qualitative investigation.

5. Recommendations

1. **Set up clear standards for data storage and dissemination within ELGO-DIMITRA, allowing for flexibility in format, storage and curation**

These could include the use of an established cluster of international databases or the setup of a centralized ELGO-DIMITRA data repository. It is crucial that the data infrastructures used or designed for this purpose can support a variety of data sharing and data ownership models, from fully open to access controlled.

The results of the survey show that ¾ of respondents would like to be able to submit their data to a repository, but the percentage is even higher when researchers are given the ability of conditional sharing. Researchers are worried both about issues of data custody and security but also about the appropriate scientific use of data. A model that allows the researchers to have some control over the access to the data, their usage as well as provides a direct peer-peer communication with the ones that ask for them, is more preferable and provides more trust and motivation for submitting data on a repository. Cybersecurity measures must also be employed early on for providing even greater trusts to researchers. Ideally, a team of specialists should be in place to maintain, curate it and improve its interface.

2. **Ensure that ELGO-DIMITRA data are easily findable, even when they are not immediately accessible**

Repositories and data platforms officially sanctioned for use by ELGO-DIMITRA should have the ability to make data findable, by signaling their presence and providing some essential metadata including how to contact original authors, without necessarily requiring researchers to make data fully accessible. This will ensure that ELGO-DIMITRA researchers can negotiate conditions of collaboration with prospective partners / data users, and gain leverage with external parties like governments and industry.

3. **Offer a training programme in data management across all institutes, with added components geared to the specific circumstances of each institute and location.**

A training programme should be devised (or adopted and adapted from available programmes from initiatives such as FAIR-IMPACT, ELIXIR or ROSiE Knowledge Hub (which is co-curated by Greek researchers) and offered to all staff. The programme should contain information and strategies for data collection, storage, management and dissemination, including indications around responsible data stewardship and which data platforms and infrastructures to use for which data types. Whether such a programme should be mandatory for all staff is to be considered: it would be useful to provide a core set of skills to all staff, though this will need to be complemented by domain-specific skills within each institute. It is recommended that such a programme be made engaging and linked tightly with the local goals and conditions of researchers within each institute, with quality tutors and good documentation made available so that researchers can continue to refer to go back to these resources long after the training is finished. Attendance and compliance with the programme should be formally acknowledged and rewarded within the ELGO-DIMITRA credit and promotion system.

4. **Establish a programme of Data Management Champions / Mentors, including appropriate rewards for the role**

To stimulate participation in training and related data management initiatives, a system to incentivize researchers who have gone through the program could be created, with more experienced researchers providing mentorship in data management to newcomers, younger staff and researchers that stumble across difficulties. A programme identifying “Data Management Champions” across the institution, with a list of experts willing and able to provide advice, would be useful. Staff identified as “champion” should be adequately recognized and compensated for the additional workload that such a mentorship task may involve. This can be achieved either through direct monetary compensation, competitions, rank promotion opportunities or any form of credit that can be desirable and valuable to researchers.

5. **Encourage the digitization of data across all institutes.**

It is necessary to improve data collection and storage practices by providing training in low-cost digital tools and their use, thereby making sure that researchers can determine whether and how the data they are generating can be digitized and made machine-readable. One quarter of the data collection practices of ELGO-DIMITRA still rely on the use of "pen and paper" technologies which are slow, fragile, non-interoperable, require lots of physical storage and management, are less transparent and cannot be part of more advanced research techniques such as asynchronous annotation, open document collaboration and querying. Not only the data needs to be digitized, but also the collection practices need to leverage affordable digital solutions for automated gathering and better precision. This involves improving knowledge not only in the use of platforms and devices but also in incorporating digital tools in research design and protocol creation.

6. **Encourage the use of data infrastructures with managed access**

Rather than encouraging researchers to share data freely online at all times, it is advisable to support the use of well-managed data platforms for data sharing, which make it possible for researchers to find which types of data are being collected and stored, and request access directly from the data producers and/or the institute locally responsible for data collection. In this way, research collaborations are being encouraged while also avoiding potential misuse of data due to inappropriate/misinformed contextualization.

7. **Set up institutional access to High Performance Computing (HPC) and provide distributed computer access to ELGO-DIMITRA researchers.**

Researchers that specifically work with modelling, remote sensing or any other specialty that deals with big data and real-time data need infrastructure to support such activities. Setting up centers for HPC with distributed compute capabilities and ability to handle an array of data formats and providing remote access to these units can vastly improve the capabilities of ELGO-DIMITRA researchers. This becomes especially important for predictive models for extreme weather phenomena caused by climate change that can lead to timely preventive measures or mitigation.

8. **Invest in better and more secure data infrastructures, either through participation in European efforts or by setting up a national structure.**

ELGO-DIMITRA can better the quality and robustness of its digital infrastructures through two main routes:

a. Taking part in EU programmes and initiatives that offer access to data center and infrastructures that ELGO-DIMITRA can't afford or doesn't have the means to maintain.
b. Actively building a new state of the art centralized national structure for use by the research institutes around Greece. ELGO-DIMITRA could either focus on building its own bespoke infrastructure or could potentially team up with universities and other research centers interested in a centralized national data service to build a commonly owned data infrastructure. This will require the full-time employment of professional technicians for maintenance and improvement of the infrastructure.

9. Encourage the move towards automation and digitalization in a purpose-specific and responsible way

It is crucial to establish conditions under which digitalization supports local research goals and context, as well as which forms of digitalization and automation of research will be beneficial and/or harmful to research practices on the ground.

10. Explore the use of novel innovative tools for decentralized storing, sharing and publishing of research data.

Given concerns about unconditional data sharing expressed by the respondents, the use of proprietary tools or publishing platforms, ELGO-DIMITRA could explore the use of innovative tools that aim to foster openness while also favoring appropriate crediting and data actionability. The Open Science and Decentralized Science (DeSci) movements has brought forth such tools as ResearchHub, Molecule's IPNFTs (Intellectual Property Non-Fungible Tokens) and DeSci Lab's Nodes among others. Some of these tools utilize blockchain technologies for immutable and public storage and querying and might require some basic understanding of such infrastructure while others less so. Forming a working group to assess such solutions and collaborate with equivalent groups across Europe will be a concrete first step towards that direction. Such an initiative would not only aid in the modernization of research infrastructure but would also give a competitive advantage to ELGO-DIMITRA institutions in the international arena since such tools are still relatively novel.

Further Reading

Arnaud E, Laporte MA, Kim S, Aubert C, Leonelli S, Cooper L, Jaiswal P, Kruseman G, Shrestha R, Buttigieg PL, Mungall C, Pietragalla J, Agbona A, Muliro J, Detras J, Hualla V, Rathore A, Das R, Dieng I, King B (2020) The Ontologies Community of Practice: An Initiative by the CGIAR Platform for Big Data in Agriculture. *Patterns* 1: 100105 https://www.cell.com/patterns/pdf/S2666-3899(20)30139-2.pdf

Beaulieu, A. and Leonelli, S. (2021) *Data and Society: A Critical Introduction*. London, UK: SAGE. https://uk.sagepub.com/en-gb/eur/data-and-society/book269709

Leonelli S, Lovell B, Fleming L, Wheeler B and Williams H. (2021) From FAIR data to fair data use: Methodological data fairness in health-related social media research. *Big Data and Society* 8 (1) https://doi.org/10.1177/20539517211010310

Leonelli, S. (2013) Global Data for Local Science: Assessing the Scale of Data Infrastructures in Biological and Biomedical Research. *BioSocieties* 8(4): 449-465. http://dx.doi.org/10.1057/biosoc.2013.23

Leonelli, S. (2017) Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology. *Data Science Journal* 16(32): 1-11. DOI: 10.5334/dsj-2017-032

Leonelli, S. (2019) Data – From Objects to Assets. *Nature* 574, 317-321. DOI: https://doi.org/10.1038/d41586-019-03062-w

Leonelli, S. (2022) Process-Sensitive Naming: Trait Descriptors and the Shifting Semantics of Plant (Data) Science. *Philosophy, Theory and Practice in Biology* 14:16 https://doi.org/10.3998/ptpbio.16039257.000000

Leonelli, S. (2023) *Philosophy of Open Science.* Elements series. Cambridge, UK: Cambridge University Press. Open Access. https://www.cambridge.org/core/elements/philosophy-of-open-science/0D049ECF635F3B676C03C6868873E406

Leonelli, S. (2024) Globalising plant knowledge beyond bioprospecting? *History of Anthropology Review*. https://histanthro.org/tag/ethnoscience/

Leonelli, S. and Tempini, N. (eds) (2020) *Data Journeys in the Sciences*. Springer. Open Access. https://www.springer.com/gp/book/9783030371760

Leonelli, S., Davey, R., Arnauld, E., Parry, G. and Bastow, R. (2017) Data Management and Best Practice in Plant Science. *Nature Plants* 3, 17086. DOI: 10.1038/nplants.2017.86

Levin, N., Leonelli, S., Weckowska, D., Castle, D., and Dupré, J. (2016) How Do Scientists Understand Openness? Exploring the Relationship between Open Science Policies and Research Practice. *Bulletin for Science and Technology Studies* 36(2): 128-141. http://dx.doi.org/10.1177/0270467616668760

Sheehan, N. Leonelli, S. and Botta, F. (2024) Unrestricted versus Regulated Open Data Governance: A Bibliometric Comparison of SARS-COV-2 Nucleotide Sequence Databases. *Data Science Journal* 19:1–30. DOI: https://doi.org/10.5334/dsj 2024-024

Williamson H, Brettschnedier J, Caccamo M, Davey R, Goble C, Kersey PJ, May S, Morris RJ, Ostler R, Pridmore T, Rawlings C, Studholme D, Tsaftaris S and Leonelli S.* (2023) Data management challenges for artificial intelligence in plant and agricultural research [version 2; peer review: 1 approved, 1 approved with reservations]. *F1000Research*, **10**:324 (https://doi.org/10.12688/f1000research.52204.2) *corresponding author

Williamson, H. and Leonelli, S. (eds.) (2022) *Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development*. Cham: Springer Open Access. https://link.springer.com/book/10.1007/978-3-031-13276-6

6. Appendix

This section contains a more detailed analysis for each of the questions (Q1-Q19) including:

- *questions text sent out to potential respondents*
- *main findings*
- *qualitative and quantitative analysis (where relevant)*
- *brief discussions*
- *raw data from free text questions*

Readers are encouraged to use the data for further analysis and comparative studies (under license CC-BY).

If you wish to have access to the full raw data files, thematic analysis and graphs, contact Fotis Tsiroukis at ft323@exeter.ac.uk

You can test the full web-based survey here: Survey Preview in Greek

0. Information and Consent

This survey is carried out by the PHIL_OS research team, in collaboration with the General Research Directorate of Agricultural Research (AGRES) for the purpose of understanding the data management habits, needs and preferences of ELGO-DIMITRA staff. It will take around 10 minutes to complete. Results will be used to inform the future research regulation within the institution, as well as the ongoing study of open research practices by the PHIL_OS team (www.opensciencestudies.eu).

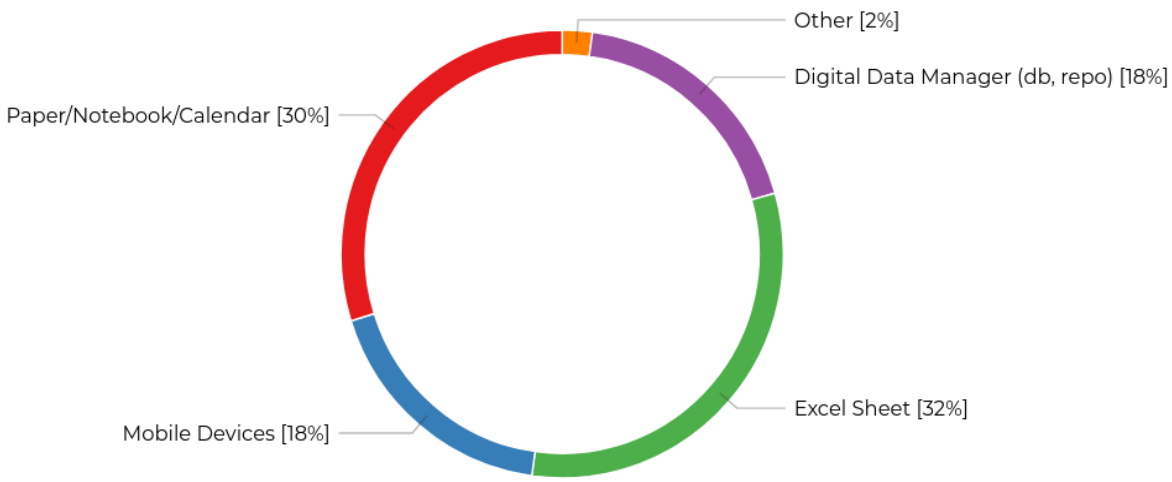
The survey is intended to be anonymous and all efforts have been made to ensure anonymity. Responses to the survey are automatically anonymized and aggregated in a manner neither allowing for the identification of individual respondents nor for the attribution of individual responses to a respondent. Although your email address may be sent along with the answers, your specific responses will not be linked to you in any way once the responses are posted online or on the AGRES website. For any question of further information, please contact Fotis Tsiroukis (ft323@exeter.ac.uk) or Sabina Leonelli (sabina.leonelli@tum.de)) or Zoi Kotsina (zkotsina@elgo.gr).

Question Text: *"Do you agree with the above terms?"*

- I agree with these terms and wish to take part in the survey
- I do not agree with these terms and do not wish to take part in the survey *(If this option is chosen, the survey logic skips to the end)*

Q1. Data Collection

Question Text: *How do you collect your research data? (pick all options that are applicable)*



Main Findings: Physical media such as paper and notebooks are quite common in the collection phase 30% of the preferred method of the respondents. The other 70% are digital media including Excel Sheets, which constitute 1/3 of the preferred method and other 1/3 being mobile devices and digital data management tools and software.

Quantitative Analysis: The question was multiple choice so participants chose more than once choice:
Respondents (Total): 92
Choices (Total): 228

Physical: [30%,68]

- Paper notes / logbook: [30%,68]

Digital: [70% 160]

- Excel Sheet: [32%,72], Mobile Devices [18%, 41], Online data manager [18%, 42], Other [2%, 5]

Discussion:

- The results reveal that the methods are fairly traditional and are that complete digitization has not taken place.
- One of the respondents answered that they collect the data through an autographic machine that instantly inputs measurement on an excel sheet. This might be important because this means that out of the respondents who chose this option, there might be more who use such equipment and that is important to contextualize the data infrastructures of the labs.

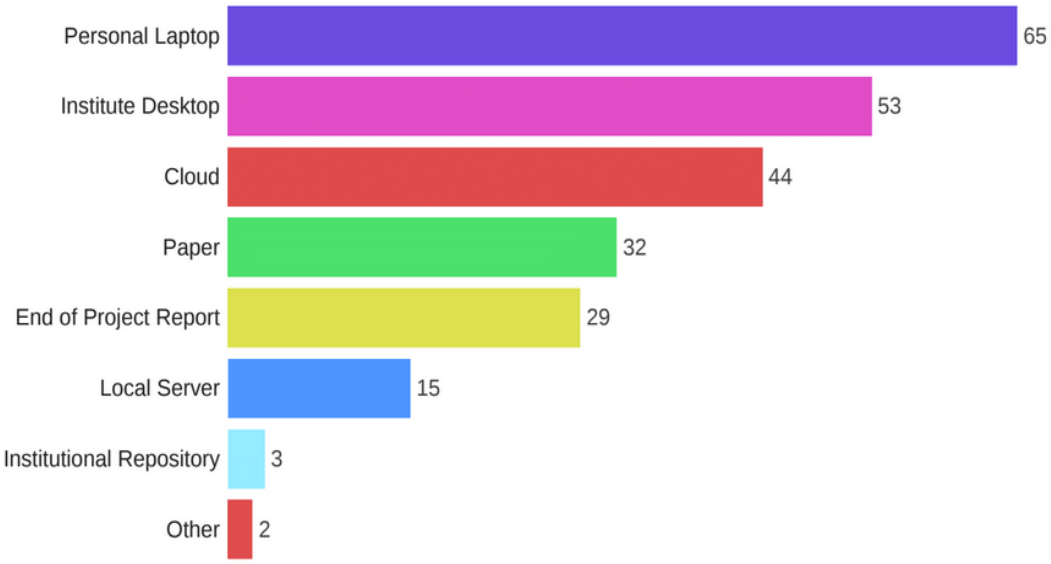
Text Entry:

Other:

- Special Protocols
- Laptop
- PC – One Drive
- Autographic Measurement Organs
- SIM card

Q2. Data Storage

Question Text: *How do you store your research data?*



Main Findings: Computer devices are the primary choice for the respondents (½ of responses), either it is a personal laptop or an institute computer. Taking into account the multiple-choice nature of the question, only a few of the respondents that chose a local computer device didn't also choose cloud and there was only one respondents that only chose cloud storage. In general most respondents choose <3 choices revealing the presence of a diversity of storage media.

Quantitative Analysis:
Respondents (Total): 91
Choices (Total): 243
Personal Laptop: 65 (26,75%)
Institute Desktop: 53 (21,81%)

Cloud: 44 (18,11%)
Paper: 32 (13,17%)
End of Project Report: 29 (11,93%)
Local Server: 15 (6,17%)
Institutional Repository: 3 (1,23%)
Other: 2 (0,82%)

An analysis was performed on the raw data to determine how many of the choices participants chose each time[1]. The results show that 51 out of 101 have >2 choices and that 20 have >3[2]. This means that a look at the graph of total responses to each choice is not enough to determine the exact use of storage technologies. A more detailed combinatorial analysis would be needed to see the exact patterns of usage between cloud, physical and local storage, but from the analysis the maximum number of choices is 5 and for all respondents who chose 5 (6 respondents in total) all include Paper and Laptop.

Discussion:

The fact that the most preferred choice is a personal laptop reveals a few important things:
a) researchers take ownership of their research project and its data and want to take care of their storage themselves.
b) stored data are more inaccessible to other researchers, more prone to loss and if the more susceptible to cybersecurity vectors if the researcher is not careful or savvy enough to take care of it themselves.
c) stored data are harder to appropriate by external researchers without crediting or to be misinterpreted.

Moreover, from the quantitative analysis of combinations of choices where ½ have chosen at least 3, it becomes clear that the data storage practices involve multiple tools and that physical and digital or local and cloud are not mutually exclusive. In fact a lot of choices include all of the above. In that sense we can interpret the graph not as showing a preference but more as revealing what is the most basic necessary equipment for data storage and what the infrastructural ecology of the everyday research includes.

In that sense, an hypothesis about the results could be that researchers tend to do a lot of work outside of the lab either in their home or in conferences and field visits where all they have available is their personal laptop device. So they mostly choose to store on a personal computer because their work involves lots of travel but also choose institutional computers when they are back in the lab. Cloud services is an intermediary solution that allows them to be able to work with both devices more fluidly. In that framing, because some answers choose both the above and also physical paper

[1] Excel Formula: =IF(A0="",0,LEN(A0)-LEN(SUBSTITUTE(A0,",",""))+1) for each individual answer.
[2] Excel Formula: =COUNTIF(An:Am, ">2")

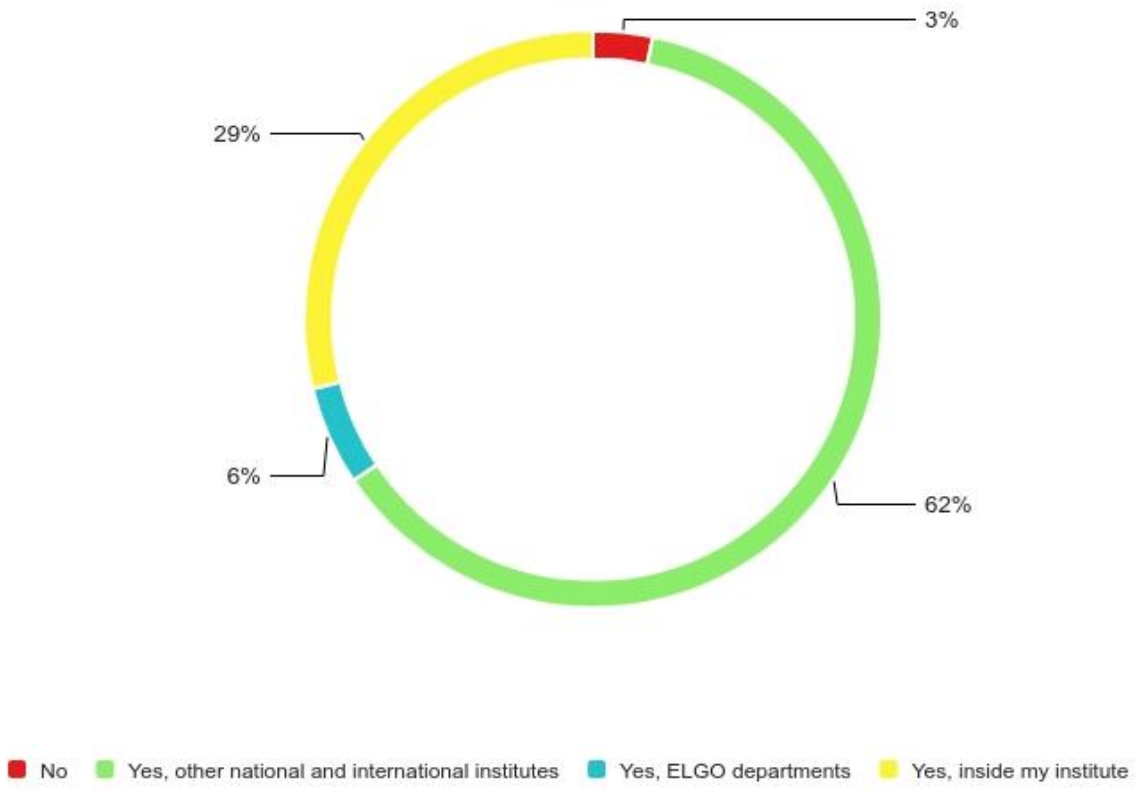
storage, we can assume that storage in paper just reveals a transition stage in storage before data are entered into a spreadsheet and stored digital or uploaded online. This means that troubling questions about the state of digitization in Greek research cannot be understood simply as pertaining to a "lack" of access to digital tools but more of a choice that reveals that there is a diversity of modes and media in data handling.
Text Entry:

Other:

- External Storage Unit
- External Hard Drive

Q3. External Sharing

Question Text: Do you share your data with researchers outside of your lab?



Main Findings: Overall 97% of researchers do share their data with other researchers outside of the lab to a certain extend. Out of these almost 2/3 (62%) choose to share

with other national and international institutes which shows that the majority chooses the greatest degree of openness to sharing out of the three options. Only 3% choose not to share beyond their own lab. Of the 97% only 6% choose to share with other HAO departments which shows that data sharing within departments inside HAO is not strong.

Quantitative Analysis:

Discussion: The implications of this outcome are quite important for understanding the culture of collaboration and communication within HAO. Sharing data that a researcher and their lab has produced can be hard, either psychologically or practically, but researchers in HAO still do so with an overwhelming consensus.

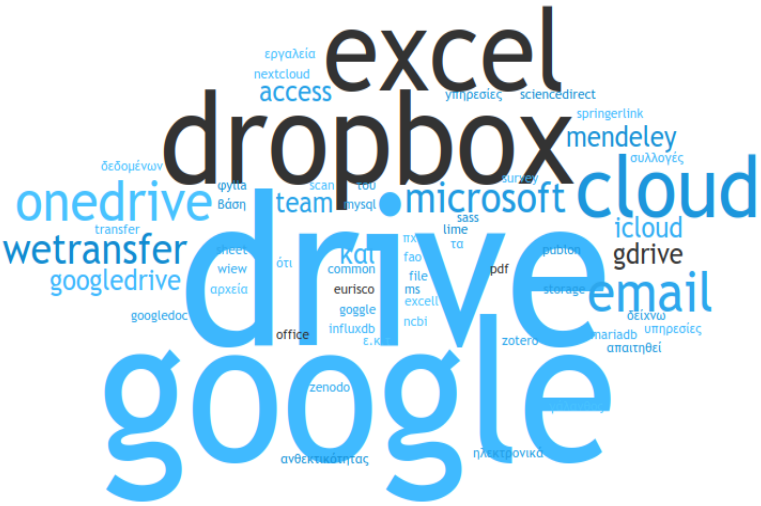
In terms of interpreting the results within the 97% that chose Yes, it might be good to clear out some potential problems in the framing of the answers. For example, in the answer *"Yes, other national and international institutes"* the scope is too big and it is not clear if people choose this because they share them mostly with other national institutes or international institutes. For example the researchers might be only sharing their research with researchers from Universities locally in their region, which reduces the scope that the question aspires too. Or they might be sharing research mostly with a neighbouring country of the Mediterranean or countries on the other side of the globe and we wouldn't know which is the case. The only thing we can infer is that the majority share data beyond ELGO-DIMITRA as an organization, which is in itself a very crucial finding.

Moreover, the fact that the researchers focus on local data sharing or data sharing outside ELGO-DIMITRA and internationally much more than inter-departmental sharing in ELGO-DIMITRA. hints to potential issues with organizational communication and maybe a lack of interdisciplinary collaboration across ELGO-DIMITRA, because if researchers prefer to share data outside ELGO-DIMITRA, the highest likelihood is that the are sharing them with researchers in close proximity to their own field and broader disciplinary communities and network of collaboration.

Q3.1 Digital Tools

Question Text: *With digital tools or databases have you used for sharing data (if they exist)?*

Main Findings: Google Drive is the dominant preference with dropbox and excel coming close. Mostly the services are proprietary cloud services most of which are freemium. A few choose a citation manager such mendeley and a small minority chooses open web platforms like zenodo.



Qualitative Analysis:

We identified 4 main categories of tools and services from respondents:

1. **Cloud Storage Services:**
    - Onedrive, Dropbox, Google Drive, iCloud, WeTransfer, Microsoft Drive, Nextcloud.
2. **Document Editors:**
    - Excel, Google Sheet, PDF Scanned Files, Electronic Files, Word Documents.
3. **Database and Research Tools:**
    - MySQL, InfluxDB, Mariadb, NCBI Database, Zenodo Zotero, Mendeley, Lime Survey, EURISCO, FAO WIEWS, SpringerLink, ScienceDirect, Publons, SASS.
1. **Email and Communication Tools:**
    - Email, Microsoft Teams

Ranking by number of choices.

1. Cloud Storage Services: 34

2. Document/File Types: 29
3. Database and Research Tools: 14
4. Email and Communication Tools: 4

Quantitative Analysis:

The top three choices with the biggest deviation from the rest are the following:

1. Google Drive (27, 20.9%)
2. Dropbox (18, 14%)
3. Excel (13, 9.3%)

The values in the parasynthesis represent the number of responses and percentage.

Ranked List of Individual Tools:

1. Google Drive (27)
2. Dropbox (18)
3. Excel (13)
4. Microsoft Teams (3)
5. PDF Scanned Files (5)
6. Electronic Files (4)
7. Word Documents (4)
8. Microsoft Office (3)
9. Email (4)
10. Whatever is required (2)
11. ZENODO (2)
12. Nextcloud (1)
13. InfluxDB (1)
14. Mariadb (1)
15. MySQL (1)
16. Lime Survey (1)
17. SASS (1)
18. WeTransfer (1)
19. iCloud (1)
20. NCBI Database (1)
21. Mendeley (1)
22. EURISCO (1)
23. FAO WIEWS (1)
24. Publons (1)
25. ScienceDirect (1)

26. SpringerLink (1)
27. Zotero (1)
28. Google Sheet (1)
29. Microsoft Drive (1)
30. Digital File (1)

Discussion: Researchers tend to choose proprietary software and closed-source apps for data sharing. From the respondents' answers the following tools are known to be open source:

1. Nextcloud
2. InfluxDB
3. Mariadb
4. MySQL
5. Lime Survey
6. ZENODO

Additionally, there are some publication platforms listed where data can be shared such as Science Direct and Publons but they are more niche choices.

In general there is great diversity of singular choices that are more specialized or niche. Observing the raw data it becomes evident that most of these were entered by only a few people who listed >2 tools. Moreover, in these cases they are thematically linked showcasing the kind of expertise of the researcher. For example, all database management tools (InfluxDB, Mariadb, MySQL) were answered by a single person (ID: R_1KetxcLT8C4tHLK) and similarly all publication platforms (SpringerLink, ScienceDirect, Publons) where similarly answered by another (ID: R_WwEPll3SIH0QJbz). This reveals that the ones who choose different platforms than the mainstream usually tend to experiment with various and also use ones that are closer to their respective specialties and research communities.

A follow-up question that can come out of these results is why do people make this choice? Is it because of familiarity, network effects, or lack of knowledge of what tools are available?

Q3.2 Refusal to Share

Question Text: Why do you choose not to share? (Text)

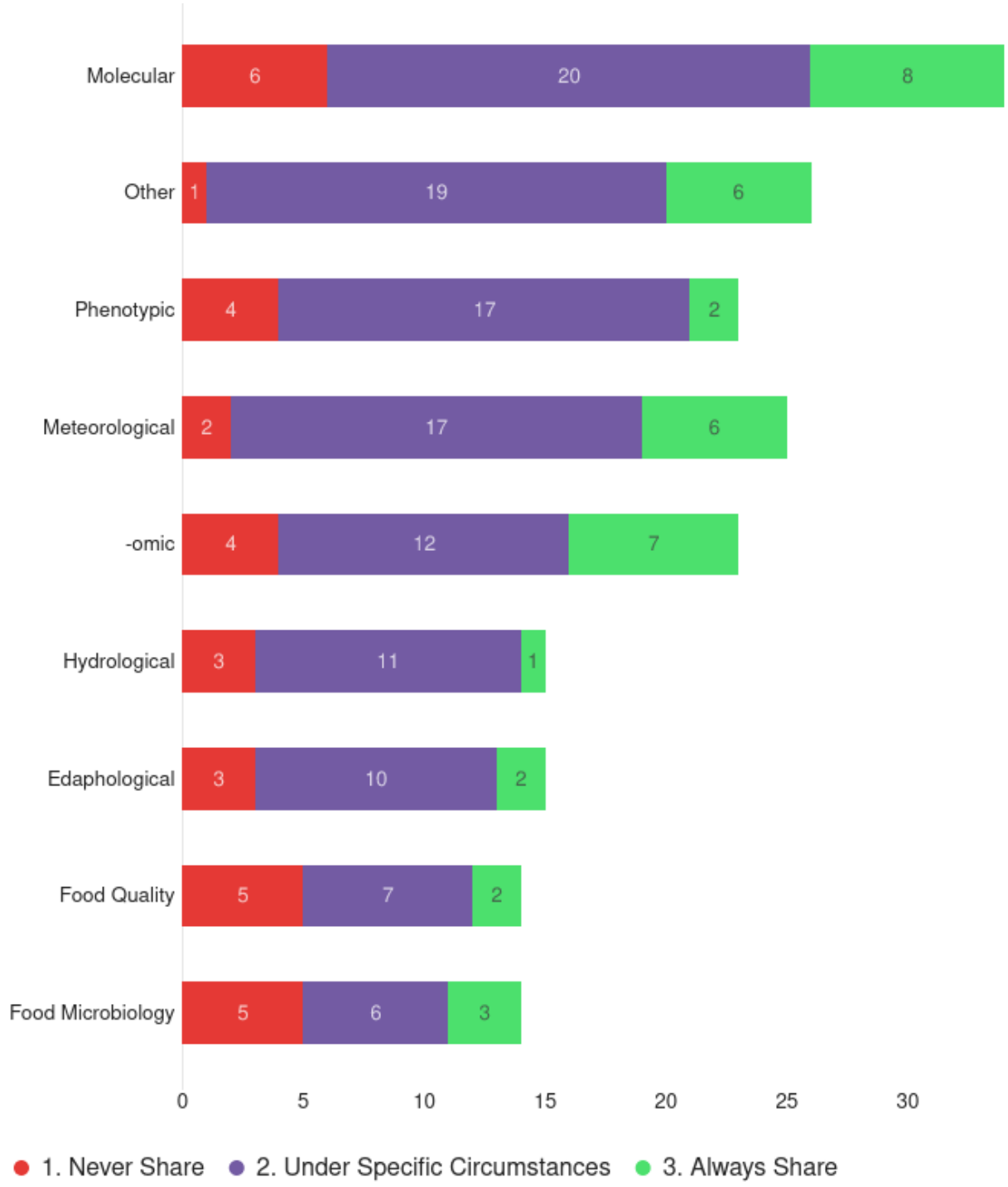
Main Findings: The ones out of the 3% in question Q3 who choose not to share answered that they share when they have been completed and published.

Discussion: The answers to this question in fact show that in fact all researchers are willing to share their data outside of their lab with the condition that they are published first. Data security and credit might be the motivators for the 3%

Q4. Most Shared Data Types

Question Text: *What types of data do you share the most?*

- Choose what type of data you produce and to what extent you share it with other researchers (inside or outside the institute).
- Fill in only the types of data you share.



Main Findings:

- Researchers share most data independent of type when they can be shared under specific conditions. In total, 63% of choices where “Under Specific Circumstances”.
- Interestingly, the second bigger choice is “Other” which, if we take into account the text answers hints at the presence of a very high degree of diversity of data types.

Qualitative Analysis:

From free text answers to “Other” we can observe a diversity of kinds of data not captured in the options of the original question. Below is a table with categorizing them into main thematic clusters and listing each individual entry. Some respondents gave a number of different entries while other a single one but for the purposes of this analysis take only each different answer into account independently of how many the respondent gave:

No.	Theme	Individual Elements
1	Agricultural Operations Data	Technical-economic data of operations, Economic data of production sectors
2	Environmental Measurements and Monitoring	Biometric data, Plant physiology measurements, bioassays, ecological data, mapping and monitoring products of habitats, environmental data, productive, Bee performance measurements, Energy, environmental, Biometrics of forest ecosystems
3	Fisheries Data	Fishing data
4	Animal Reproduction Data	Data on the fertility of productive animals and their gametes
5	Social and Behavioral Data	Data on behavior and opinions of respondents, Social and demographic data, Economic and sociological data, Educational and informational material
6	Plant Physiology	Plant physiology measurements
7	Plant Protection	Plant protection
8	Acoustic Data	Acoustic files, Sound data
9	Miscellaneous Data	Data collection and evaluation

Because of the diversity and number of responses, there are number of ways of clustering the data.

Quantitative Analysis:

Free Text answers to *“Other”*:

- Total Individual Responses: 28
- Total Respondents: 23
- Single-Element Responses: 7

Discussion: The variety of data types that respondents added in the “Other” section is a very important piece of data providing an insights in the vast diversity of data types shared by researchers. It also gives a hint as to the big range of interdisciplinarity of the whole institute and the need to have effective ways of interdisciplinary coordination and data integration across domains.

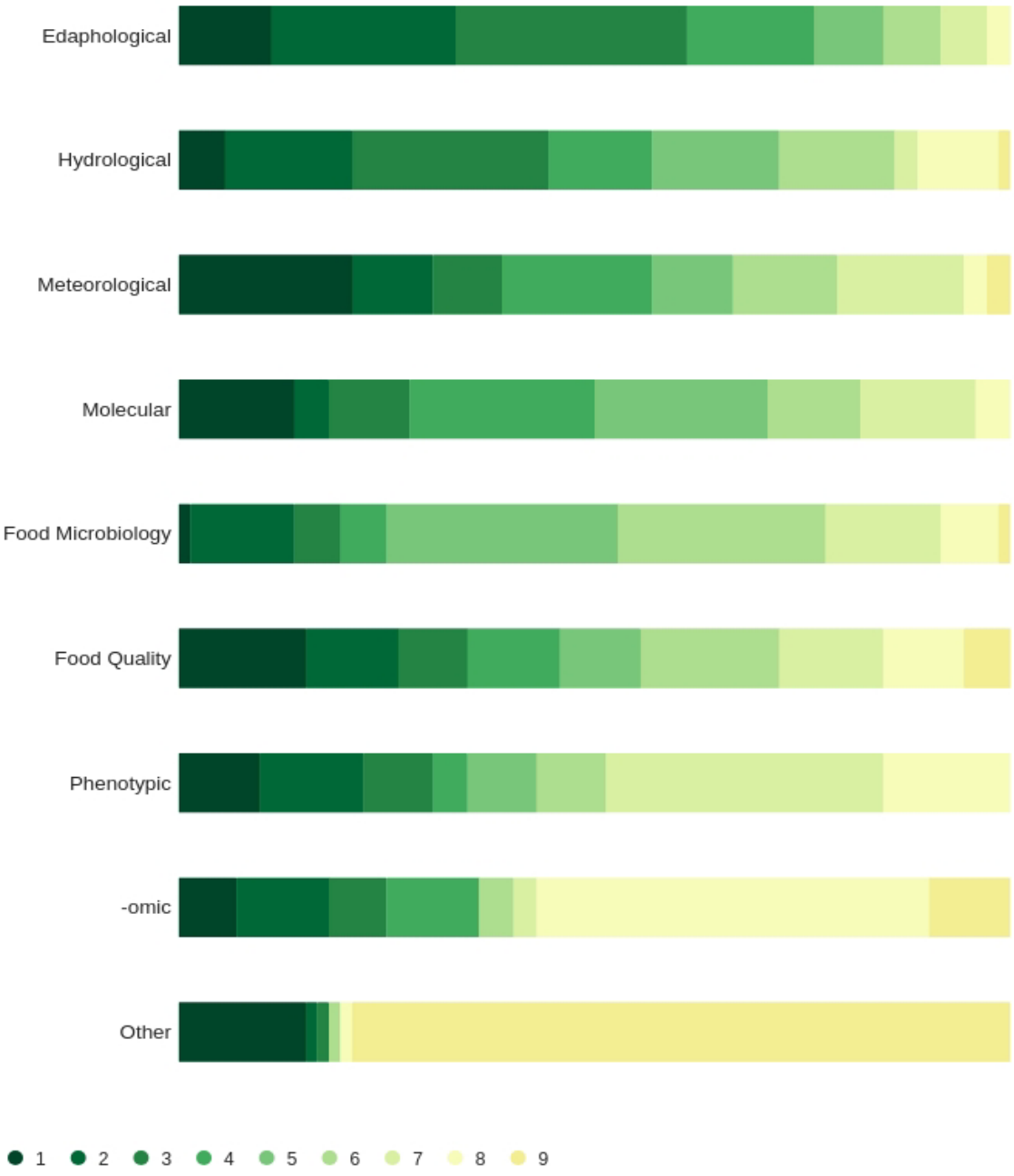
Q5. Societal Impact of Data Types

Question Text: *What kinds of data do you think, if shared outside ELGO, would be the most useful for science and society?*

> *Rank the options from most useful to least useful.*
>
> *Notes:*
> *- If on a computer, drag the options with your mouse to put them in order.*
> *- Otherwise, if using a mobile phone, drag and drop the options using the touch screen.*

Main Findings: Meterological are the data type the comes up as first the most and Food Microbiology data the least. The ones that choose other, rank it as first. But in general, relative preferences about the societal impact of data seems to be varied.

Quantitative Analysis:
Other: 16 responses



Discussion: Interpreting the results from this question is not an easy task. There are no significant patterns that can be inferred from the visualization. But this is due to the fact that the question and the format of the answers is most properly framed as supportive to other questions (especially Q4) and is more informative when looking deeper at the data.

Q6. Preferred Data Storage

Question Text: *Where do you think your research data should be stored*.

Main Findings: There are 4 dominant media for data storage: Cloud, Server, Databases and Personal Computers.

Quantitative Analysis: Because there were a lot of short repeated words in the raw data, keyword analysis was performed on them (after they were translated form Greek to English) so as to scope the frequency and percentage of the main choices. Answers that contained variations that were semantically related to the main keywords are counted as instances of the keyword (e. g. "cloud service" => "Cloud"). Here are the relative percentages for each keyword:

1. Cloud (23) - 26.74%
2. Server (19) - 22.09%
3. Database/Databases (15) - 17.44%
4. PC/Computer/Personal Computer (15) - 17.44%
5. ELGO-DIMITRA (4) - 4.65%
6. Digital (4) - 4.65%
7. Institutional Repository (3) - 3.49%
8. Backup (3) - 3.49%
9. Access (3) - 3.49%
10. Hard Disk (2) - 2.33%
11. Controlled (2) - 2.33%
12. Local (2) - 2.33%
13. Platform (2) - 2.33%
14. Electronic (1) - 1.16%
15. Security (1) - 1.16%
16. Open (1) - 1.16%
17. In-Situ (1) - 1.16%

Qualitative Analysis: The keyword frequency analysis gives a very good idea of the main patterns present in the dataset. But it only shows the frequency of singular words while some individual answers in the dataset included combinations that show a relationships between them. This is much more common in the "Server" answers which tend to qualify the kind of server storage. Here are a few examples:

> In servers with controlled access (7), In cloud services or servers within ELGO-DIMITRA institutes (40), On servers of the Institute and/or the Research Directorate of ELGO with free access to the services of the Ministry of Agriculture and Universities, based on reciprocity (56).

Text Entry:

All Responses (translated from Greek to English):

1. Cloud
2. Cloud, common storage in PCs
3. Clouds
4. On computers or servers that meet security conditions
5. In relevant databases
6. Local computers/servers, external drives, and cloud platforms
7. In servers with controlled access
8. Computer
9. In my own digital media
10. Cloud + Hard Disk
11. Server
12. Cloud
13. Hard copies and electronic copies
14. Database in the cloud
15. Server
16. PC, Cloud
17. Server
18. In institutional repository
19. Locked in the cloud
20. Server automatically and in the cloud
21. Cloud
22. Personal computer
23. On my computer servers (ELGO-DIMITRA)
24. Database of the organization
25. On the computer
26. In a platform with controlled access for accredited users
27. Personal computer or Institute's data bank (Laboratory Information System), which I acquired but has not been installed for objective reasons yet
28. On my computer and in the institutional repository
29. In databases on the internet
30. In open databases
31. Cloud
32. In cloud services
33. In digital databases
34. In central repositories of large organizations
35. Databases
36. Cloud
37. Cloud
38. On the server of the unit
39. Digital with the ability to access from another computer

40. In cloud services or servers within ELGO-DIMITRA institutes
41. Exclusively on servers belonging to ELGO-DIMITRA, protected from cyber attacks, with backups
42. In a multitude of files (e.g., spreadsheets), including files provided by companies manufacturing research measurement instruments
43. Cloud service
44. Server
45. Cloud
46. Servers
47. Cloud
48. In electronic platforms
49. On a personal computer and on a server
50. On the local server of the Institute
51. In a repository within the Institute, in addition to storing on a personal hard drive
52. Personal platform
53. There could be a digital library with limited access to maintain the data
54. Server
55. In the cloud and simultaneously on physical storage units
56. On servers of the Institute and/or the Research Directorate of ELGO with free access to the services of the Ministry of Agriculture and Universities, based on reciprocity
57. Institute's database
58. Cloud
59. On the server of ELGO-DIMITRA
60. Cloud
61. Institute's server
62. Digital
63. In-Situ and Cloud
64. On some institutional server or cloud
65. On a personal computer/cloud
66. In a secure place
67. PCs
68. In databases with backups (server, cloud, etc.)

Q7. Challenges

Question Text: What challenges do you face in relation to accessing and handling research data?

Main Findings: Most researchers face problems with infrastructures related to storage. The common themes are volume, speed and lack of backups. An interesting pattern in the answers was that some respondents linked their problem to a specific data type (e.g. storage of acoustic data).

Discussion: An interesting thread is the one concerning interoperability between data formats, data centers and standards for sharing and storing different forms of data, as well as the issue of data aggregation and the difficulty of access that comes with the non-existence of aggregation platforms.

Text Entry:

1. None
2. Internet connection speed
3. None
4. None in particular
5. Verification is time-consuming and costly because it is done on infrastructure not belonging to ELOG-DIMTRA.
6. Large volume of acoustic data. Takes a long time even for backup. Currently sharing through postal shipments of hard drives.
7. Data is not centralized.
8. Often work with large time series that need to be sent telemetrically, and there is no good quality network for their transfer.
9. REDUCED STORAGE CAPACITY
10. Internet speed
11. Not all data are available.
12. Limited storage space and difficulty in transferring large files.
13. Categorization
14. Network speed
15. Large volume
16. We don't have our own SERVERS.
17. Storage space (referring to my own data) and corresponding management and analysis tools
18. Access to others' data is difficult and sometimes requires evaluation due to incorrect estimation (determination) of their values.
19. No access to servers.
20. Insufficient backup, problematic file management on different computers.
21. No access to data from other laboratories.
22. None
23. NONE
24. I don't know what data are available.
25. Difficulty in access
26. Nothing specific

27. Annual subscription payments
28. Acquisition of an environment for storing and sharing data
29. Insufficient storage space
30. No databases on the internet providing access to research data, such as meteorological, hydrological, etc.
31. Lack of training and information provision regarding open databases.
32. No information storage and processing system for large data.
33. As the volume of data increases, management becomes more difficult.
34. The volume of files is now very large.
35. Lack of technological tools and equipment for easier entry and access to research data.
36. Platforms are not always user-friendly.
37. None. When I don't have access, I ask for them.
38. Lack of software
39. Need for digitization
40. File size, availability of specialized algorithms for data analysis, computing power.
41. Organization of a large volume of data
42. Limited storage space using commercial means (e.g., Google Drive) and not enough computing power.
43. They are scattered in multiple locations (Dropbox, Google Drive, iCloud, Nextcloud, etc.). Difficulty in use and sharing with colleagues, as well as a significant personal financial cost.
44. Changes in operating systems, incompatibility due to software updates.
45. Heterogeneity of storage types among colleagues, requiring multiple storage in different databases.
46. Not shared by colleagues, and it is uncertain if other colleagues will appropriate them.
47. No central storage point for all data, and therefore, no direct access.
48. Lack of space
49. AVAILABILITY OF STORAGE SPACE
50. DELAYS
51. None.
52. Access to data stored on servers within the Institute, from computers outside the Institute.
53. In general, there is no uniformity of rules for accessing data (availability, cost, etc.). There are data (such as topographic, coverage, and land use data, meteorological) that are required as infrastructure for analyses and should be provided freely by the state.
54. Availability of a suitable platform for storing/managing a database.
55. There is not always access to them.
56. Existence and training in new statistical analysis data packages.
57. Right of access
58. Existing databases do not have a unified architecture and are not always accessible.
59. Open data located in the Ministry of Rural Development & Food database.
60. Management of data volume - fast servers
61. No access or subscription is not paid on time.
62. NONE
    -
63. DIGITAL
64. Very large volume of data (acoustic) often amounting to 1-2 TB per project, making storage and sharing difficult.

Q8. Support

Question Text: What kind of support would be most useful to help you manage and analyse your data?

Main Findings: Among the most frequent themes are support (with IT, statistics, specialist analysis), education and software-related issues (cost, licenses, support, updates). As a secondary cluster respondents reported issues with storage infrastructures, such as lack of local and institutional servers, database storage platforms and equipment.

Discussion: Same as in Q7, an interesting pattern in the answers, especially when it comes to support and calls for bringing in experts, is the discipline-specificity or data-type specificity of the responses. Most have to do with lack of knowledge or support with statistics and IT/databases but there are also interesting cases of expertise needed such as bio-informatics and econometrics.

Text Entry:

1. I don't need support.
2. I am covered.
3. From specialized analysts, e.g., econometrics.
4. Provision of equipment and consumables.
5. Access to free or very cheap online storage for TB of data (tens).
6. Provision of statistical programs or specialized data analysis programs.
7. Contracts with providers for telemetric transfer and storage of data through the ELGO.
8. Access to data analysis software and training (similar to universities).
9. Server manager.
10. Statistical processing.
11. A database and a cloud.
12. I would like to know the total data available and to which I can have access.
13. Infrastructure of a local server and network speed.
14. Workshops.
15. Facilities.
16. Mainly software (SAS, ARGGIS, MindManager, AI-based applications for data analysis).
17. I don't know.
18. Technical support for information systems of the institute.
19. Unlimited cloud space and a local server.
20. Portable electronic devices.
21. None.
22. (Empty line)
23. To have ELGO-DIMITRA servers with sufficient computing power.
24. Support for data analysis with new software.
25. (Empty line)
26. Purchase of software, subscriptions.
27. IT expert.
28. Any support is welcome.
29. Additional work is required for processing and storing data before they are stored in databases.
30. Training from research institutes and providers of databases (Google, Microsoft, EU/ERA).
31. Creation of an information system for data collection, storage, and processing, and for providing information and knowledge.

32. Updates on management programs.
33. Statistical analysis.
34. Technological equipment, software licenses, training seminars.
35. Specialized personnel.
36. Software, hardware.
37. Provision of suitable data analysis tools, e.g., SYSTAT, MATLAB, etc.
38. Provision of statistical packages.
39. Greater computing power (access to a cluster), open analytical tools.
40. Support for the method of data analysis (e.g., statistical analysis) possibly through seminars, for data management (writing method), and analysis.
41. Specialized IT services in each institute and centrally.
42. Central IT and simultaneously IT in each institute that offers immediate solutions to problems that arise.
43. Support for the software provided by companies with the current operating systems.
44. Personalized storage/management service in the cloud.
45. Someone data analyst.
46. Cloud service for storage and data analysis software (e.g., statistical packages, visualization).
47. IT.
48. Purchase of software.
49. Training.
50. Strengthening of human resources.
51. Technical support for remote access to these and software licenses for processing and analysis of data.
52. To have a large metadata database, searchable with keywords and information on where they are available and under what conditions.
53. Platform management seminar.
54. Technical and updates are always useful in statistical analysis packages.
55. Education and provision of statistical programs.
56. Specialist in organizing databases.
57. Server.
58. Server within ELGO-DIMITRA and per institute.
59. I don't know.
60. Specialist in bioinformatics.
61. Digital.
62. Access to a fast cloud database for TB of data - this, however, requires our institute's network to have a corresponding connection.
63. Tools and trained personnel that ELGO does not have.
64. Face-to-face seminars.
65. Training seminars.
66. Employment in each institute of technical information systems and suitable equipment and training.

Q9. Old Data

Question Text: Do you have problems accessing old information, such as data or metadata collected more than 10 years ago? If so, please indicate specific cases.

Main Findings:

Quantitative Analysis:

1. No: 21 times (35%)
2. Yes: 14 times (23.3%)

3. Meteorological data: 10 times (16.7%)
4. Printed format: 2 times (3.3%)
5. Not digitized: 4 times (6.7%)
6. Difficulty: 2 times (3.3%)
7. Accessible: 3 times (5%)
8. Grapevine varieties: 1 time (1.7%)
9. Phenotyping: 1 time (1.7%)
10. Traceability: 1 time (1.7%)
11. Institution's management: 1 time (1.7%)
12. Private files: 1 time (1.7%)
13. Archived books: 1 time (1.7%)
14. Electronic file: 1 time (1.7%)
15. Hard drives: 2 times (3.3%)
16. Hard copy: 2 times (3.3%)
17. Outdated software: 1 time (1.7%)
18. Incompatible: 1 time (1.7%)
19. Personal files: 1 time (1.7%)
20. Hellenic National Meteorological Service: 1 time (1.7%)
21. Public Power Corporation: 1 time (1.7%)
22. Ministry of Environment and Energy: 1 time (1.7%)
23. Hellenic Statistical Authority: 1 time (1.7%)
24. ELGO-DIMITRA: 2 times (3.3%)
25. Experimental data: 1 time (1.7%)
26. Evaluation of varieties: 1 time (1.7%)
27. Institute: 1 time (1.7%)
28. Non-digitized data: 2 times (3.3%)
29. Previous research: 1 time (1.7%)
30. Notebooks: 1 time (1.7%)
31. DVDs: 1 time (1.7%)
32. Minimal data loss: 1 time (1.7%)
33. Recovered: 1 time (1.7%)
34. Forest fires: 1 time (1.7%)
35. Paper/article: 1 time (1.7%)
36. Incorrect classification: 1 time (1.7%)
37. Old computer: 1 time (1.7%)
38. External hard drive: 1 time (1.7%)

Note: Percentages are rounded to the nearest tenth.

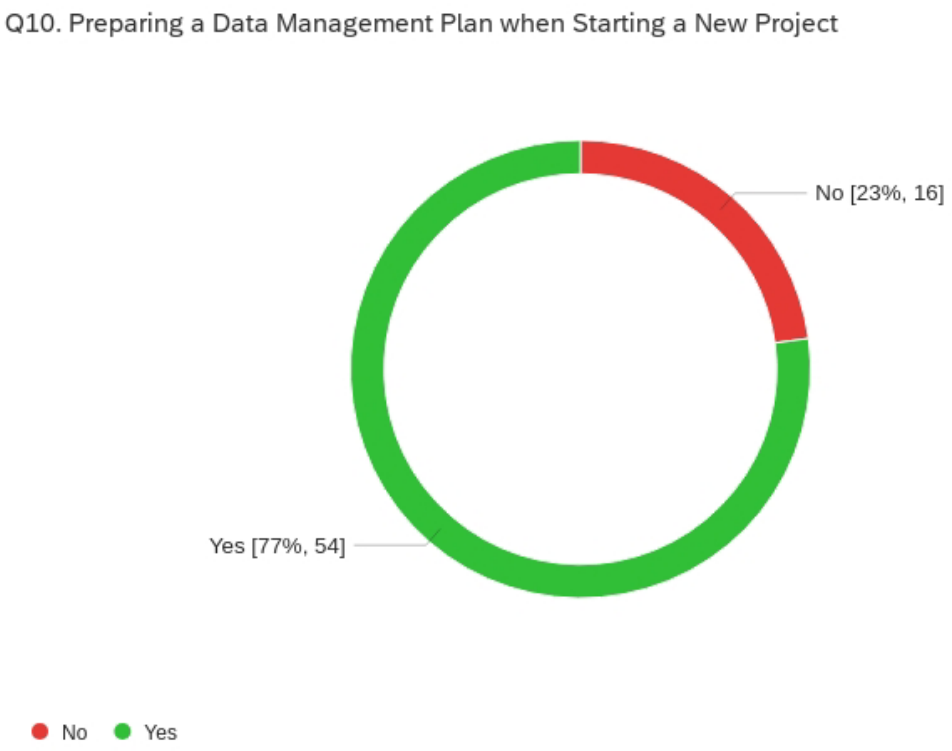
Discussion:

Text Entry:

1. No.
2. No.
3. Yes, on the internet.
4. No.
5. Meteorological data from previous years in printed format, not digital.
6. The files have not been digitized.
7. Not difficult as long as they exist to some extent on the internet and are freely available for use. Difficulty exists for those that are not accessible.
8. No.
9. No.
10. Yes, meteorological data.
11. No.
12. Yes. Mainly concerns phenotyping data for grapevine varieties and their traceability. Many of them are not in the institution's management but in the possession of private files of the institution's employees.
13. No.
14. Yes. They are all in archived books.
15. No.
16. No.
17. Yes, because the information is not in an electronic file.
18. Stored on various hard drives and/or in hard copy format.
19. Yes, when the data are in files or folders.
20. No.
21. Yes, due to outdated software that no longer works, and the stored database cannot be opened by new software since they are not compatible.
22. Only through personal files.
23. No.
24. Yes, such as meteorological data from the Hellenic National Meteorological Service, Public Power Corporation, and the Ministry of Environment and Energy.
25. No.
26. Yes, in meteorological and data from the Hellenic Statistical Authority.
27. There are no phenotypic data for very old varieties of ELGO-DIMITRA. There are no older experimental data from the evaluation of varieties by ELGO-DIMITRA at the Institute.
28. No.
29. No.
30. No.
31. Non-digitized data from previous research. However, the data existed in a file with access.
32. Yes, in cases where they are recorded in notebooks/notebooks/documents.
33. Usually, this data is on hard drives or even DVDs that are now difficult to access or cannot be read.
34. Minimal data loss. Usually recovered.
35. No.
36. No access at all.
37. No, at least not currently, processing such data.
38. In many cases, it is difficult to find detailed information about past forest fires.
39. No.
40. What was not published in the form of a paper/article, we do not know where both the primary data and any conclusions from their processing are.
41. The problem arises from incorrect classification of data in files.
42. If they exist, they exist only on an old computer or an external hard drive.
43. No.
44. No.
45. No access to such data.
46. No.
47. Meteorological data.
48. Yes, data in printed form that have not been digitized.

Q10. Project DMP

Question Text: *Do you prepare a data management plan when starting a new project?*



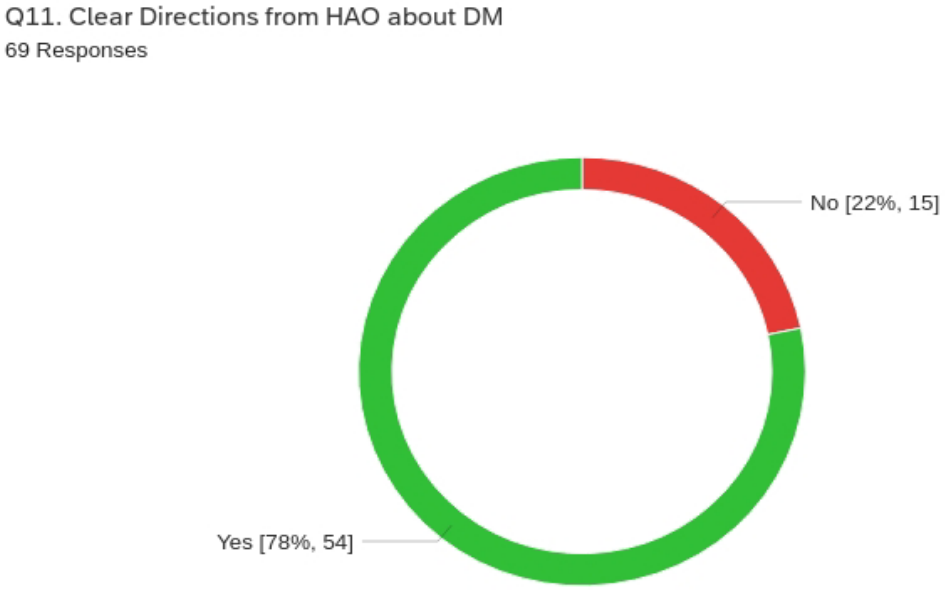
Main Findings: More than ¾ of respondents prepare a DMP before starting a new project.

Quantitative Analysis:
Total Responses: 70
Yes: 54 (77%)
No: 16 (23%)

Discussion: Overall the results show that DMP is a significant part of the research practices of ELGO-DIMITRA researchers and that there is awareness about its importance. But the number of negative answers is big enough to signal that some of the research might struggle from issues of lack of organization of research data which also impacts the transparency of the research. A 23% is big enough to act as a motivation for ELGO-DIMITRA to improve its educational and incentive structures to support data management.

Q11. ELGO-DIMITRA Guidelines

Question Text: *Would you like to receive clear guidelines from ELGO-DIMITRA related to ways of managing your research data?*



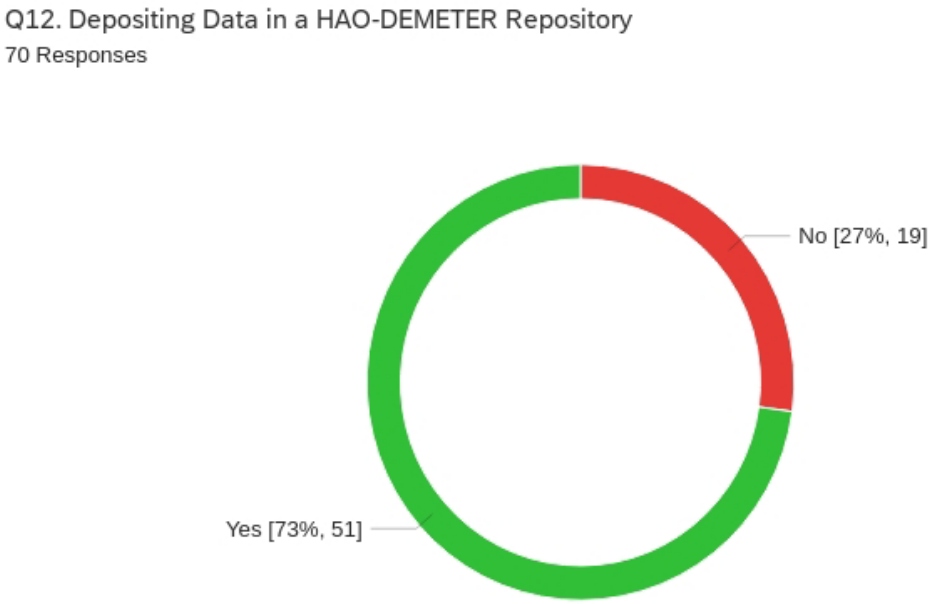
Main Findings: More than ¾ of respondents would like to receive clear guidelines about managing research data from ELGO-DIMITRA.

Quantitative Analysis:
Total Responses: 69
Yes: 54 (78%)
No: 15 (22%)

Discussion: In general, the results show it would be a good idea for ELGO-DIMITRA to provide clearer guidelines and documentation of various ways of data management. What is not clear from the way the question was phrased is whether the respondents expect this to in respect to ELGO-DIMITRA centralized repositories and platforms or whether the researchers expect from ELGO-DIMITRA to provide discipline-specific guidelines.

Q12. Data Repository

Question Text: *Would you like to deposit your data in a ELGO-DIMITRA data repository?*



Main Findings: Around ¾ of respondents would deposit their data on an institutional ELGO-DIMITRA repository. This is a significant majority but it is still far from consensus.

Quantitative Analysis:
Total Responses: 70
Yes: 51 (73%)
No: 19 (27%)

Cross-comparison with Q13:
Yes: 61 (87%)
→ +10 Respondents, +14%
→ Out of 10: Q12|No / Q13|Yes = 5, +7%
No: 9 (13%)

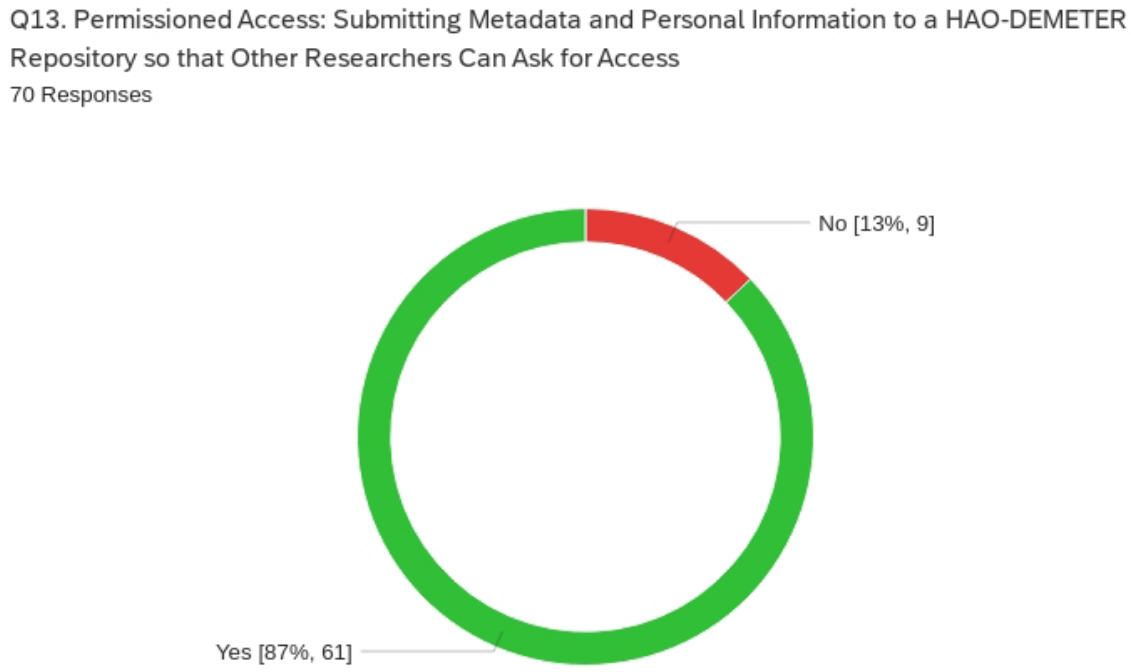
Discussion: Assessing the meaning of the results is better done when they are compared to the results of the other Yes/No questions. In the comparisons a few insights becomes clear:

- Q12 is the question with the least amount of “Yes” answers compared to the other 5 (Q10-Q14), even if the significant majority is still affirmative of the question.
- When compared with Q13, which has 14% more “Yes” answers, it is clear that a condition for submitting data is that there is more control on their access form

the side of the researcher who produced them as well as more information about them. Looking at the raw data (HAO_Survey_Results_Raw.csv/.xlsx) also reveals that some of the researchers who answered "No" on Q12, answered "Yes" on Q13 (5 respondents, 7%) which provides the two conditions above to researchers.

Q13. Metadata Submission

Question Text: *Would you be willing to submit information about your data (and not the data itself) and your details to a ELGO-DIMITRA data repository, so that researchers interested in your data can contact you for access?*



Main Findings: Researchers are more likely to submit their data on a repository with a permission access model rather than submit their data in an unconditionally open way.

Quantitative Analysis:
Total Responses: 70
Yes: 61 (87%)
No: 9 (13%)

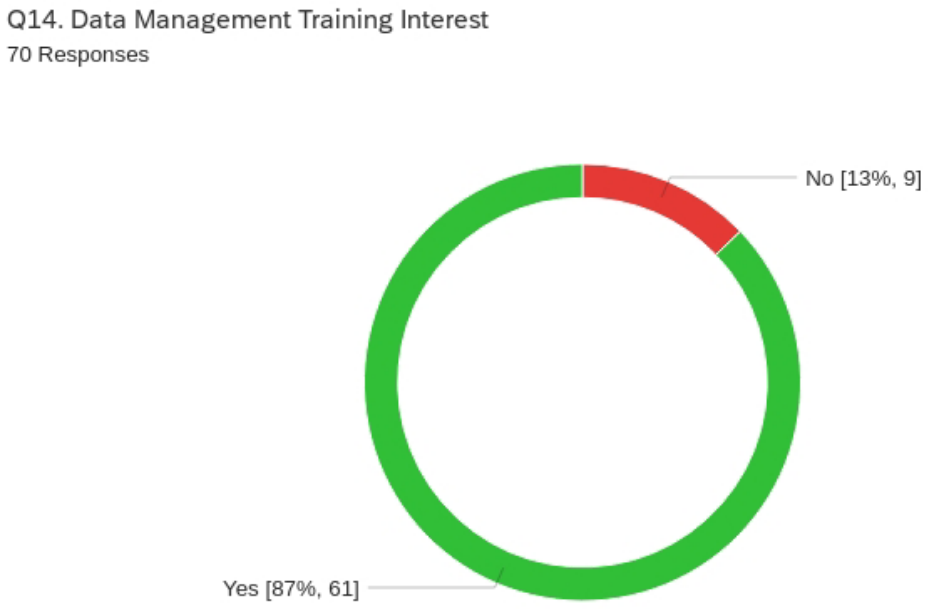
Discussion: Same as in Q12 plus it is important to mention that there are no respondents who answered "Yes" on Q12 that didn't also answer "Yes" on Q13. This

reveals that conditional sharing in an institutional repository is 14% more preferable to unconditional sharing.

As mentioned in the previous question, the conditional nature of uploaded data is important for providing confidence and trust to researchers that their data will be used in the right way. It is important to mention however that, the 7% that answered No to Q12 but Yes in Q13 is not that big of a difference. This means that the majority of researchers are already willing to more open forms of data sharing on an institutional repository and that providing more control and information is a factor that provides more trust and security to the ones that would be already willing to share.

Q14. Training

Question Text: Would you be interested in being trained in data management?



Main Findings: A significant majority of respondents prefers to be trained in data management.

Quantitative Analysis:
Total Responses: 70
Yes: 61 (87%)
No: 9 (13%)

Discussion: The percentage of affirmative answers is enough to signal to ELGO-DIMITRA that kick-starting an initiative for running training courses for researchers in HAO institutes is a priority. The results can be interpreted in two important ways:

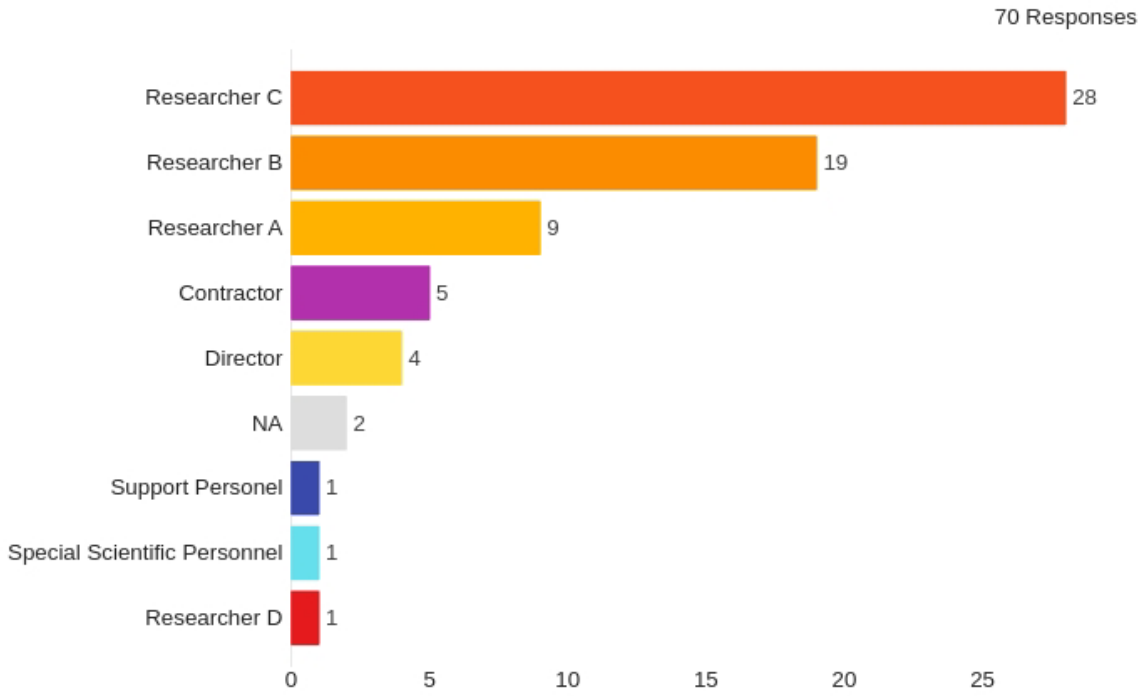
a) Showing an interest in learning more about data management because of insufficient knowledge or for gaining more confidence and skill. This could be because of a broader lack of data management education, especially in Greek higher education.
b) Good data management practice is recognized as an important aspect of research and there is an expectation from the broader research culture for researchers to have good data management practices.

What lies behind the results is most likely a combination of learning and expectations and this can show the way towards more precise follow up questions such as:

- Do you think there is lack of sufficient data management education?
  - Is there a difference between Greek universities and ones abroad?
- Do you feel that there is external pressure and expectations for good data management?

Q15. Role

Question Text: *What is your role in ELGO-DIMITRA?*



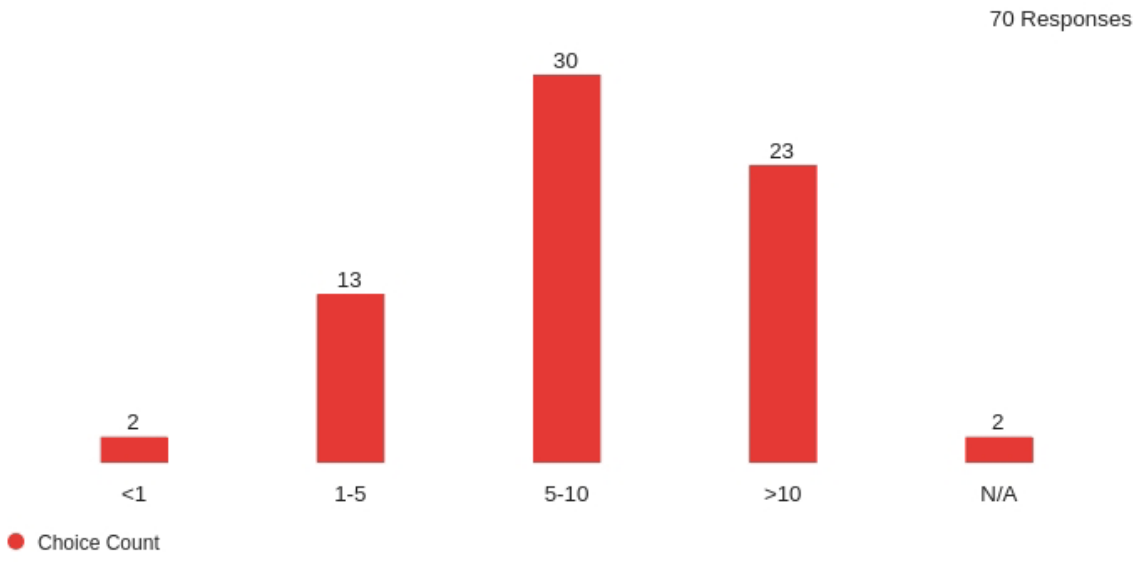
Main Findings: Most of the respondents are middle-tier researchers in ranks B and C. However, only one researcher was in Rank D while we would expect more researchers to be in that rank.

Quantitative Analysis:

Discussion:
Total Responses: 70
Director:
A: 9 (11%)
B: 19 (28%)
C: 28 (40%)
D: 1 (0,7%)
Contractor: 5 (7%)
Support Personnel: 1
Special Scientific Personnel: 1
N/A:

Q16. Years of Employment

Question Text: *How many years have you been working at ELGO-DIMITRA?*



Main Findings: The majority of respondents has been in ELGO-DIMITRA for >5 years and 1/3 has been employed for >10 years.

Quantitative Analysis:

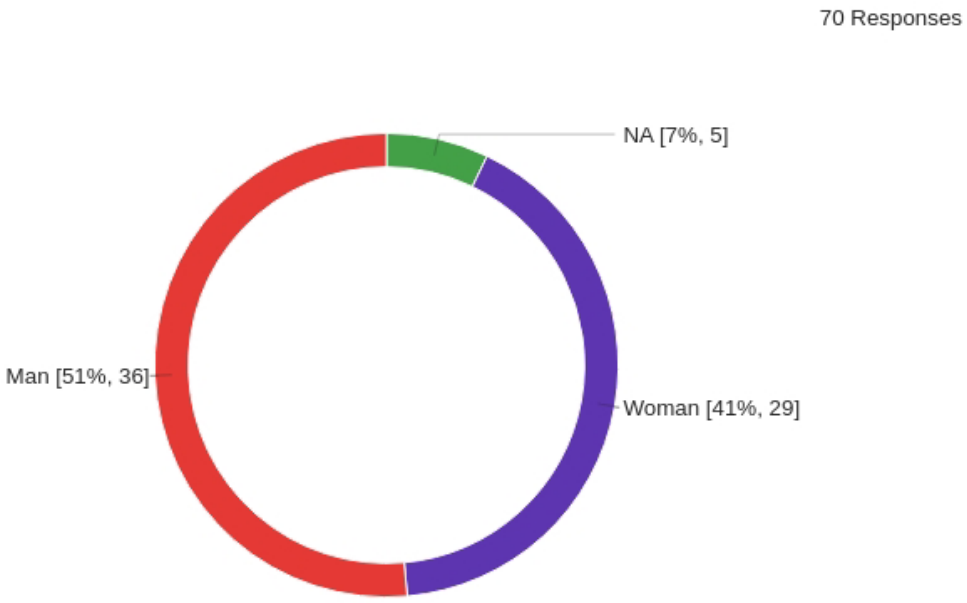
Discussion: What the data results mean is that that people who work at ELGO-DIMITRA tend to be in permanent positions and that there are benefits in staying at a post for more than a decade. This is a good thing for the forging of good and effective

collaborative relationships in departments and for seeking long-term research projects that could bring lots of fund and resources to the institutes.

These results are consistent with the results of Q15, where most of the respondents seem to be at Ranks B and C, which require a career investment of more than 5 years to reach.

Q17. Gender

Question Text: *What is your gender?*



Main Findings: The overall gender distribution looks fairly even with a relatively small disparity of 10% between men and women.

Quantitative Analysis:
Total Responses: 70
Man: 36 (51%)
Woman: 29 (41%)
N/A: 5 (7%)

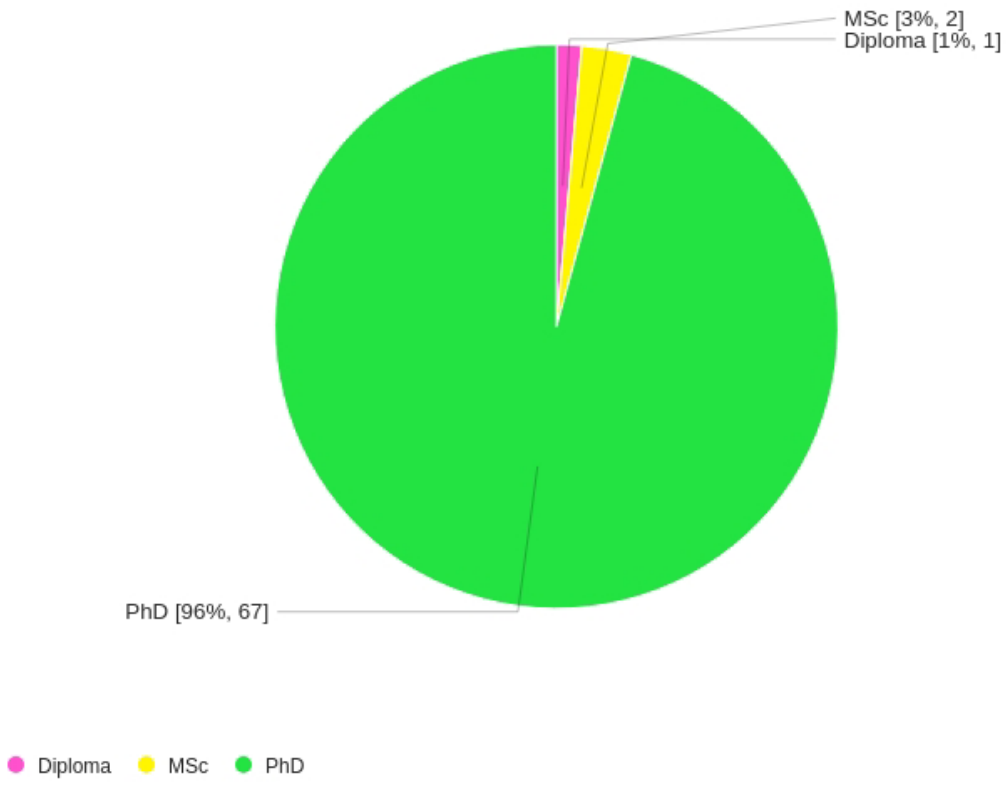
Cross-comparative:
Contractors: 1/5 men, 4/5 women
Non-PhDs: 0 men

Discussion: The graph shows a relatively equal gender distribution. But the graph from this question by itself can be misleading unless we look deeper into data of individual respondents across questions. A cross-comparative look at the raw data from Q15-19 shows that all but one of the employees in contractor roles are women. Moreover, the workers at other supportive and auxiliary roles as well as on the D rank, are all women. Additionally, cross-comparing with Q18, shows that all non-PhDs are also women (MSc graduates and diploma holders). These are patterns in the data that must be taken into account when we attempt to interpret the results of this question in terms of categories about equal gender distribution or equity,

Προσωπικό υποστήριξης	Επαγγελματικό Δίπλωμα	Γυναίκα
Προσωπικό με σύμβαση	Μεταπτυχιακός τίτλος	Γυναίκα
Προσωπικό με σύμβαση	Διδακτορικό δίπλωμα	Γυναίκα
Προσωπικό με σύμβαση	Διδακτορικό δίπλωμα	Γυναίκα
Προσωπικό με σύμβαση	Διδακτορικό δίπλωμα	Γυναίκα
Προσωπικό με σύμβαση	Διδακτορικό δίπλωμα	Άνδρας
Ειδικό Επιστημονικό Προσωπικό	Μεταπτυχιακός τίτλος	Γυναίκα
Ερευνητής/τρια Δ	Διδακτορικό δίπλωμα	Γυναίκα

Q18. Qualifications

Question Text: *What is your highest qualification?*



Main Findings: Almost all respondents are PhD graduates. Only 4% have a lower qualifications. There were none with undergraduate degrees and none that chose "Other".

Quantitative Analysis:
Total Responses: 70
PhD: 67 (96%)
M.Sc: 2 (3%)
B.Sc: 0
Diploma: 1 (1%)
Other: 0
N/A: 0

Discussion: The results show the profound lack of students and younger researchers in ELGO-DIMITRA. Not being a university research environment it makes sense for most employees to be highly qualified.

But this finding might potentially show the lack of diversity in positions available and the lack of infrastructures for supporting new graduates in agriculture.

Although the presence of PhD shows that the requirements expected from new employees are quite high and show signal the quality of ELGO-DIMITRA as a research environment, it also shows that the voices represented might be skewed towards the academic world.

The biggest drawback potentially is that deep experiential knowledge of research practices from senior researchers that takes bigger time to fully incorporate is missing from ELGO-DIMITRA. This might signal a lack of knowledge-transfer from the older generation of agricultural researchers to the younger. A suggestion is therefore the addition of studentships, training and mentorship programs for students and young researchers.

However, because of their contractual nature, a great deal of these workers might have not been represented in the responses received, either because directors might have not shared the survey with them, or because the contractors might have not considered themselves as part of the ELGO-DIMITRA human resources.

Q19. Comments

Question Text: *Add any other comments or thoughts you wish (optional).*

Main Findings: The comments provide a more granular insights into more specific issues and suggestions that researchers have such as:

- The importance of *incentive structures* for data sharing.
- The need of strict criteria to avoid “theft” and “bad usage” of research data.
- Need for *clarification* of processes.
- Need for *digitization* of documents
- Enabling *Inhouse analysis and storage* of data through the setup of servers and units with sufficient computational power.
- DMP training should be mandatory since DMPs are a prerequisite in international projects.
- Data *cybersecurity* and anonymity.
- Need for institutional availability of genetic material data.