Data literacy and data research management: results from a Portuguese survey among researchers and academics

Ana Lúcia Terra[1], Ana Alice Batista[2], Carla Teixeira Lopes[3], Cristina Ribeiro[3], Fernanda Martins[4], Gabriel David[3], Irene Rodrigues[5], José Borbinha[6], Maria Manuel Borges[7], Maria Manuela Pinto[4], Paulo Fialho[8]

[1] Polytechnic Institute of Porto. CEOS.PP | CIC.Digital, Porto, Portugal
anaterra@iscap.ipp.pt
[2] ALGORITMI Center, University of Minho, Portugal
analice@dsi.uminho.pt
[3] INESC TEC, University of Porto, Portugal
{ctl, mcr, gtd}@fe.up.pt
[4] CIC.DIGITAL Porto, University of Porto, Portugal
{mmartins, mmpinto}@letras.up.pt
[5] University of Évora, Portugal
ipr@uevora.pt
[6] INESC-ID, IST, Lisbon University, Portugal
jlb@tecnico.ulisboa.pt
[7] University of Coimbra, Portugal
mmb@fl.uc.pt
[8] IVAR, University of Azores, Portugal
fialho.paulo@gmail.com

Abstract. This study reports the Portuguese contribution to an international survey on data literacy of academics and researchers are presented in this study. The community contributed with 943 filled questionnaire, covering key aspects related to the use and production of research data (e.g. file type and volume of data created and used; the choice of data storage devices and the creation of metadata on research data, among others). Also considered were the use of Data Management Plan and data management practices (e.g. file naming, citation rules, use of unique identifiers and tags), as also sharing of research data. Based on the results, it is concluded that there is a need to formulate institutional policies for the management of scientific data and to design training initiatives to develop data literacy skills. The comparing of these results with those of the overall international study is a next step.

Keywords: Data literacy, data research management, scientific community, Portugal.

1 Introduction

In the last decades, the amount and variety of data produced by researchers has created what Borgman [1] dubbed the data deluge. The availability and volume of data with business and scientific relevance has grown dramatically, both in the public and in the private sector. However, in order for data to become available to both the scientific and business activities, it is necessary to make them accessible, intelligible, assessable and usable. The premises of Big Data, eScience and open access to research data require data management strategies to promote data sources and data maximum use. This framework led some governments and supranational institutions to develop guidelines for the management of scientific data in order to enhance their broad access. The European Union has identified two problems in open access to research data: a lack of coherent open data ecosystem and a lack of attention to the specificity of research practice, processes and data collection [2].

In this context, concerns about the creation, access, organization, sharing and preservation of research data have gained significant visibility, emerging the concept of data literacy. According to Calzada Prado & Marzal [3], data literacy concerns the capabilities of individuals to access, interpret, critically assess, manage, handle and ethically use data, presenting a strong connection with information literacy. Koltay (2015a) also stresses the relationship between data literacy and information literacy, in particular as regards the skills covered. Based on information literacy skills formulated by information literacy standards, Carlson, Fosmire, Miller, & Nelson [4] present a set of core competencies for data literacy structured around 12 themes. The content of each theme should be adapted to the specificities of different scientific areas. The topics covered include: 1. İntroduction to databases and data formats, 2. Discovery and acquisition of data, 3. Data management and organization, 4. Data conversion and interoperability, 5. Quality assurance, 6. Metadata, 7. Data curation and re-use, 8. Culture of practice, 9. Data preservation, 10. Data analysis, 11. Data visualization and 12. Ethics, including citation of data. This framework is intended to develop data management skills for data-producing and data-using researchers, from the dual perspective of researcher-as-producer and researcher-as-consumer. The data literacy skills were also analysed by Calzada Prado & Marzal [3] who defined a framework with five core competencies, covering 10 topics. This data literacy framework is geared towards sustaining training programs. Skills and teaching topics include: 1. Understanding data (1.1. What is data?, 1.2. Data in society: a tool for knowledge and innovation), 2. Finding and/or obtaining data (2.1. Data sources, 2.2. Obtaining data), 3. Reading, interpreting and evaluating data (3.1. Reading and interpreting data, 3.2. Evaluation data), 4. Managing data (4.1. Data and metadata collection and management), 5. Using data (5.1. Data handling, 5.2. Producing elements for data synthesis, 5.3. Ethical use of data).

Given the skills involved in these two models, a wide range of information practices can be integrated in information literacy field. For the purpose of this paper, we may consider, the definition of the Association of College and Research Libraries [5] "Information literacy is the set of integrated abilities encompassing the reflective discovery of information, the understanding of how information is produced and valued, and the use of information in creating new knowledge and participating

ethically in communities of learning".
According to Koltay [6], research data management and data quality are directly dependent on data governance and data literacy and should be based on research data services. Data governance can be defined as a system that distinguishes decision and accountability profiles for processes related to information management, establishing procedures for who can execute what actions, based on what information, when, under what circumstances and following which methods [7]. By encompassing individual's skills to access, interpret, critically evaluate, manage, manipulate and use the data, data literacy allows the transformation of data into information and from this into applied knowledge. In addition, data literacy is directly aligned with the scientific methodology of any area by fostering methodological transparency, data preservation, data sharing and re-use as well as accountability.

2 Methodology and sample

The Portuguese results of the international survey about data literacy from academics and researchers are presented in this paper. The Data Literacy and Research Data Management project was developed by an international team from UK, France and Turkey, in order to collect data about data literacy of academics, researchers and research students in higher education institutions. In a second phase, the project was open to include other countries, to allow a broader comparative study. To facilitate cross-country comparisons, a common questionnaire, using English language, was created by the initial team and then translated in the language of each participating country. The Portuguese version was disseminated at the beginning of January 2017 and contributions were received until April. The Portuguese team includes researchers from universities and polytechnic institutes, the two higher education sub-systems of the country. Each researcher was responsible for the survey distribution in his/her institution, thus creating a convenience sample with academics, researchers, and research students.
The questionnaire included 26 questions organized in two groups. The main findings of the two groups will be presented and analysed. The first group aims to collect demographic information, including occupation, age, discipline, gender, country and institution. The second group focuses on awareness of data management issues, including aspects such as data type and volume used by researchers, data sources, produced data types and volume, data storage devices, and metadata addition to research data. Issues related to data sharing and data storage are also included. Awareness with respect to data management and to the processes associated with research data are also assessed, namely concerning the use of data management plans, at the institutional or individual level, and the use of metadata associated with research data. This was covered on questions about standard file naming system, DOI and ORCID identifiers, guidelines for citing data, or data annotation. Training on issues concerning research data management, such as data management plans, metadata, consistent file naming or data citation styles, were also surveyed.
The Portuguese survey was started 1.946 times, and 943 complete surveys were collected. The data analysis will be based the total number of completed surveys. The

first response was recorded on the 10th January 2017 and the last on the 16th April 2017. Regarding the demographic characterization [Q4], the male gender (54.83%) is predominant. As for the age range [Q2], three groups stand out with close percentages: 26-35 years (23.65%), 36-45 years (28.53%) and 46-55 years (29.80%). The 56-65 years' group has a percentage of only 11.66%. Thus, the sample includes for the most part adults with age between 26 and 55 years. In terms of research experience [Q5], 27.47% of the respondents are researchers with more than 20 years of experience and 25.66% between 5-10 years. 14.21% has dedicated 11-15 years to research and 14.74% dedicated 16-20 years. Thus, the sample has a significant research experience, with only 16.33% having less than five years of experience.

The most represented scientific areas [Q3] are Engineering and Technology (27.89%) and Natural Sciences (22.80%). On the Engineering and Technology group, the prevailing sub areas are Electrical engineering, electronic engineering and information engineering with 16.86%. The social sciences compose 18.77% of the sample and humanities composes 10.60%. The sample includes 9.23% individuals from medical and health science and agricultural science has a residual percentage of 1.80%. 8.91% respondents chose the "other" option, however when asked to specify it, they were framed within the given options.

3 Results

3.1 Research data creation and use

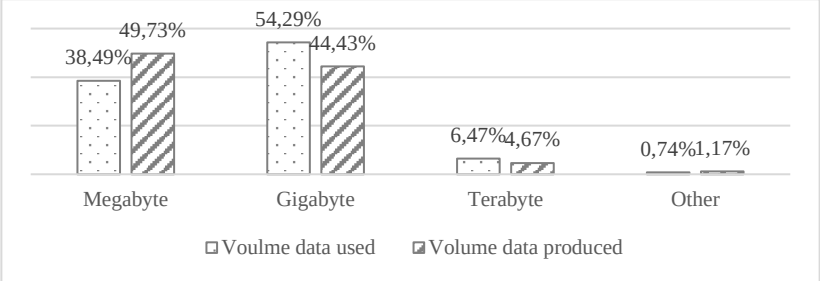
The type of files used and created in the scope of research are wide-ranging [Table 1].

Table 1. Data file type used and created for research [Q8, Q12]

	File type you use	File type you produce
Standard office documents	87,17%	80,81%
Structured scientific and statistical data	54,19%	50,37%
Encoded text	23,75%	17,71%
Internet and web-based data	54,29%	19,83%
Databases	37,75%	22,27%
Images	59,60%	44,54%
Audio files	18,66%	8,06%
Structured graphics	10,39%	9,86%
Raw (machine-generated) data	26,94%	14,32%
Archived data	51,33%	23,01%
Software applications	40,51%	20,68%
Source code	27,36%	19,19%
Configuration data	25,56%	13,36%
Non-digital data	36,37%	21,00%
Other	4,45%	2,65%

The standard office documents option, including text, spreadsheets and presentations represents 87.1% of the files used and 80.81% of the files produced. These types of files are a basic support for not only data acquisition, but also for research production, mainly research dissemination. Actually, to write papers, reports and working papers, researchers typically use these files. In this sense, it is questionable that almost 20% of the respondents did not choose this option regarding the research data creation. A little more than half of the sample also uses structured scientific and statistical data, e.g. SPSS, GIS, (54.19%) and the internet and web-based data, webpages, e-mails, blogs, social network data, (54.29%). File images (JPEG, GIF, TIFF, PNG, among others) are also used by a significant number of respondents (59.60%) as well as archived data (e.g.: ZIP, RAR, ZAR), option chosen by 51.33% of the respondents. It is also noteworthy that over a third of the respondents (36.37%) state not to use digital data and 21% state not to produce digital data. As a result of their investigation, 50.37% state to create structured scientific and statistical data, 44.54% images, 23.01% archived data, 22.27% databases, 20.68% software applications and 19.19% source code. These last percentages should be related with the fact that over a little more than a quarter of the sample represents the Engineering and Technology area. As for the Table 1, it is also noteworthy that the sample shows a bigger range of file type used than file type produced.

Fig. 1. Volume of data used and produced for research [Q9, Q13]



The data volume is connected to the range of type of file used and produced during the research process. According to Figure 1, over half the sample (54.29%) uses gigabytes of data, but only 44.43% produces data of the same order of magnitude. Actually, the megabyte is prominent as the sample data production (49.73%). On the other hand, the terabyte was selected by a residual percentage of the sample: 6.47% for used data and 4.67% for produced data. Based on this chart, one can conclude that the quantity of data used is larger than the data produced.

3.2 Storage and Processing research data

The organization of research data, to make then usable, is an important component of data management. For this purpose, it is necessary to attend to data storage, data identification and data processing, especially when they are from outside sources.
Data storage options are crucial because they uphold data access and preservation in the short and long term. Almost all respondents claimed (94.38%) to use their own devices

(e.g.: computer, tablet, external drive) to store their data [Q14]. Half indicated to also use a cloud service (50.80%). Only 30.65% chose the central servers/repositories of the university option and 12.83% selected outside repositories. One can conclude that each researcher/research team seems to manage their data storage independently, according to their specific context conditions. So, it doesn't seem to exist an institutional infrastructure of data storage, where backup and storage solutions are defined in a short and long term. Metadata creation about research data also constitutes an important component of data management. Respondents were asked to indicate what type of additional information they assign to their research data [Q15]. Attribution of administrative information (e.g. creator, date of creation, file name, access terms/restrictions) was referred by 45.92%. Discovery information (e.g. creator, funding body, project title, project ID, keywords) has almost the same percentage (44.64%). Technical information creation, e.g. file format, file size, software/hardware needed to use the data (26.62%), and description of the data file, e.g. file/data structure, field tags/descriptions, application rules (29.59%), present considerable lower percentages. 26.41% states that no additional information is assigned to their research data. After research data gathering, one important step to enable its use is related to the previous work of structuring and organizing which will allow its manipulation and analysis. Having this in mind, respondents were asked on how they use external data sources [Q11]. Over half (55.46%) stated that it took a bit of effort for some cleaning and/or modifications. 30.22% refer spending a lot of time and efforts to make it usable for the project; and 24.60% claim to use the data as collected.

3.3 Research data management

A Data Management Plan (DMP), the use of metadata, using a file naming convention, using citation rules, using DOI and ORCID, applying rules on data storage or promote data sharing are important issues for research data management. As such, respondents were asked to comment on these aspects [Table 2].

Table 2. Issues about research data management (Q19)

	Yes	**Uncertain**	**No**
Does your institution have a DMP?	3,61%	77,31%	19,09%
Have you ever used a DMP for your research?	4,88%	14,95%	80,17%
Do you have a DMP for your current research project(s)?	5,83%	17,82%	76,35%
Do you think a DMP actually helps researchers in managing research data?	31,71%	60,13%	8,17%
Are you familiar with the term metadata?	61,19%	11,13%	27,68%
Do you think a formal training on metadata would be useful for managing research data?	60,45%	34,46%	5,09%

	Yes	Uncertain	No
Does your university have a prescribed metadata set for uploading data to a repository?	15,16%	76,14%	8,70%
Does your research community use/recommend any standard file naming system?	11,45%	47,19%	41,36%
Does your university have a standard/consistent file naming system?	6,26%	64,37%	29,37%
Do you use any standard style for citing research data?	62,35%	9,33%	28,31%
Are you familiar with the concept of DOI?	69,88%	7,42%	22,69%
Does your university recommend any specific guideline for citing data?	42,52%	37,54%	19,94%
Have you got any unique researcher identification (like ORCID)?	77,84%	7,32%	14,85%
Does your university actively encourage you to share data on open access mode?	25,56%	48,25%	26,19%
Are you familiar with your university and/or funding body's requirements with regard to data storage?	16,86%	29,37%	53,76%

With regard to DMP, respondents reveal little familiarity with the concept. In fact, 77.31% indicate that it is uncertain whether their institution has a DMP. Besides, 80.17% indicate they have never used a DMP and 76.35% do not have a DPM for their current research projects. The usefulness of this document is clear only for 31.71% of the sample, with 60.13% stating to be uncertain about this matter.

The concept of metadata seems clearer for the sample because 61.19% indicate to be familiar with it and 60.45% thinks that formal training on metadata would be useful for managing research data, although more than a third is uncertain about this. The fact that 76.14% is not sure about the fact that his/her university has a prescribed metadata set for uploading data to a repository is also worth mentioning.

Regarding the use of a standard/consistent file naming system by their research community, respondents showed uncertainty (47.19%) or stated that this is not true (41.36%), with only 11.45% stating that this practice exists. In the more global context of their university, only 6.26% indicate there is a standard/consistent file naming system, with 64.37% being uncertain about this matter.

Regarding the use of a standard style for citing research data, 62.35% indicate to do so. However, in the more general context of the university, the recommendation to use a specific style (e.g. APA, Harvard) is not very common. In fact, only 42.52% state that this recommendation exists in their university.

The use of unique identifiers for informational objects and individuals is an increasingly important element in the scientific data management. The majority (69.88%) is aware of this and claims to be familiar with the Digital Object Identifier (DOI) concept. This is reinforced by the fact that 77.84% have an unique researcher identification (e.g. ORCID). However, 22.69% indicate that they are not familiar with DOI and 14.85% do not have an unique researcher identification.

As for their knowledge of whether their university encourages data sharing in open

access, 48.25% indicates to be uncertain. Additionally, only a quarter (25.56%) state that their university encourages data sharing in open access and 26.19% indicates that this does not happen. This lack of knowledge about the policies of their institution is also visible when respondents indicate that they are unfamiliar with the requirements of their university or research funding agency regarding data storage requirements. In fact, 53.76% indicate that they are not familiar with the subject.

3.4 Research data share

Regarding their data share practices [Q16], respondents state to share their data with researchers at the same team (72.64%) and with researchers in other institutions (50.48%). However, 10.82% does not share data and only 34.78% shares their data with researcher in the same university.

Additionally, they were also questioned regarding the type of access they allowed to their data [Q17]. Conditioned access in several levels prevails: 38.39% allows open access to their research data for their team members; 40.72% grants access upon request and 29.06% claims that their data have restricted access (e.g. only some parts of the dataset is accessible). Less than a fifth (18.98%) grants open access to their data, and on the other hand, 9.12% claims that their data is not available to anyone else.

To understand aspects that can limit data share, respondents were asked to comment on some statements [Q18]. Thus, 27.15% feels no concern about sharing their data. The concerns that mostly limit data sharing are of legal and ethical order (45.39%), as well as the feeling of lack of appropriate policies and rights protection (22.48%). The misuse of data (45.28%) and the misinterpretation of data (29.59%) constitute two other obstacles to data sharing. It is noteworthy that fear of losing the scientific edge (17.18%) and lack of technical, financial and personnel resources (7.23%) are less significant. Thus, according to the respondents, the inhibitions to data sharing do not relate to a closure position in relation to the rest of the scientific community, in order to safeguard some personal advantage, but rather due to external issues such as the absence of a legal framework that protects research data. In addition, improper use of the data, outside the context and the objectives for which it was collected, also makes it difficult to share data. The sharing of research data was also addressed in Q21, where respondents had to comment on some statements related to this subject, among others. The majority of the respondents (57.47%) stated that they felt comfortable and willing to share their research data with others, and there seemed to be a favourable attitude towards this sharing. In addition, 45.81% do not foresee problems in sharing their data. However, it should be emphasized that 84.31% perceive data ethics could be an issue when research data is shared with others. In this regard, it seems necessary to create a regulatory framework and practical conditions to address this concern in order to encourage research data sharing. These assumptions are part of the open access requirements for research data, with 52.28% declaring familiarity with this concept.

3.5 Data management practices

In order to understand the data management practices, one of the questions listed a set of related tasks asking to indicate the frequency in which each task was performed [Table 4].

Table 4. Frequency in which each data management task is performed (Q20)

	Almost Always	Often	Sometimes	Rarely	Never
Using metadata standard for tagging your data	3,29%	7,21%	15,27%	22,80%	51,43%
Using your own/in-house (your research team) tags and metadata	9,86%	17,82%	21,74%	14,10%	36,48%
Using datasets that are tagged with standard metadata	2,86%	9,33%	20,78%	20,15%	46,87%
Using file naming convention or standard	11,24%	16,54%	18,56%	17,50%	36,16%
Having different versions of the same dataset(s)	17,82%	27,89%	28,00%	14,21%	12,09%
Using systems/techniques for version control to easily recognise a specific version	19,62%	21,42%	18,66%	15,16%	25,13%
Citing research data	41,04%	30,22%	16,65%	6,57%	5,51%
Working with data that are generally in the public domain	18,56%	33,40%	26,09%	14,95%	7,00%
Working with data that have restricted access?	10,39%	25,45%	27,25%	22,38%	14,53%

The option with the highest percentage was never using metadata standard for tagging your data (51.43%). There are also some high percentages in the options never using datasets that are tagged with standard metadata (46.87%) and never using your own/in house (your research team) tags and metadata (36.48%). Thus, tagging practices, which help organize data by identifying their content in a structured and normalized way, are not common. In fact, there is a high respondent's percentage who states to have never tagged and the group who make it regularly (almost always and often) are a minority.

Another relevant practice for digital research data management is using rules for file naming and version control. Only 27.78% states to regularly use (almost always and often) file naming convention or standard, and 36.16% claims to never have done it. Handling the gathered data and its subsequent configurations leads to the creation of different versions. This is a usual reality (almost always and often) for 45.71% of the respondents, and only 12.09% claim it to be otherwise. As such, version control is an important strategy to correctly analyse data and it is of common use (almost always and often) for 41.04% of the respondents. None the less, 25.13% of the respondents do not have that practice and 33.82% does it only casually (sometimes and rarely).

The correct use of research data implies its citation, which is a common practice (almost always and often) for 71.26% of the respondents. However, almost a quarter (28.73%) states never doing it or not regularly (sometimes and rarely). Regarding the type of data, half of the respondents (51.96%) states to use almost always or often data from the public domain, and 35.84% uses data from restricted access.

4. Discussion

Regarding the type of files produced and used in the research, it stands out that more than a third (36.37%) state not to use digital data as an information source. Almost a quarter (21%) indicates not to produce digital data. In a context where the use of ICT is ubiquitous, this data is somewhat surprising. It would be important to determine whether this non-use is based on the fact that the digital data is not at all necessary for the research that these researchers carry out or due to a lack of digital skills. In this case, it seems relevant to develop data literacy skills at the level of data creation and use, elements of various data literacy training programs [3, 4].

The surveyed scientific community seems to manage the collected research data independently, without supervision framing of professional affiliation institutions or research funding institutions, since almost everyone indicates that they use personal devices to store the data. In this context, the creation of institutional data storage infrastructures could facilitate the work of researchers, ensuring research data preservation in medium and long term. The creation of a DMP at institutional level could support this approach, but it appears that there is only a residual percentage (3.61%) of respondents stating that their institution has this document. In addition, 80.17% indicate that they have never used a DMP in their research. At institutional level, therefore, there is a need to create technological infrastructures to support the management of scientific data, as well as to develop policy guidance documents and to design standardized scientific data management practices.

Although this institutional framework/support is lacking, respondents say they use practices that help them manage and work on research data. Thus, the addition of metadata related to administrative information, discovery information, technical information or tags is performed by varied percentages of respondents, even if they are not preponderant habits. Here too, there seems to be a very broad field of specialized training needs on the part of the scientific community surveyed.

Another aspect that can be underlined is the fact that 77.84% of respondents declare using unique researcher identification, such as ORCID. This practice has been recommended or required by higher education institutions and research funding agencies. Therefore, the existence of a formal framework demonstrates practical effects in terms of changing researchers' behaviour. In this sense, many of the data management practices can be improved if an institutional framework exists.

5. Conclusions

This research produced comprehensive knowledge about the data literacy and data research management practices of a significant and diversified sample, in the Portuguese context. The most relevant data was presented but its analysis could be deepened by verifying the disparities of knowledge and practices among researchers from different scientific areas. Another relevant comparative approach for a better knowledge of the Portuguese scientific community may be to compare the results of the affiliated respondents to the university sub-system and to the polytechnic sub-system.

The results presented and discussed were obtained from the application of a web-based questionnaire, with the respondents self-declaring behaviours and knowledge. In this

sense, the level of knowledge and practices reported may present some disparities regarding the actual skills of the sample. However, the data collected supports the need to promote training initiatives aimed at developing data literacy skills, framed in the formulation of research data policies.

References

[1] Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and*

[2] *RECODE: Policy recommendations for open access to research data*. (2014). Retrieved from

[3] Calzada Prado, J., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs : core competencies and contents. *Libri*, *63*(2), 123–134.

[4] Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: a study of students and research faculty. *Portal: Libraries and the Academy*, *11*(2), 629–657.

[5] ACRL. (2016). *Framework for information literacy for higher education*.

[6] Koltay, T. (2016). Data governance, data literacy and the management of data quality. *IFLA Journal*, *42*(4), 303–312.

[7] The Data Governance Institute. (2017). Definitions of Data Governance. Retrieved May 11, 2017, from http://www.datagovernance.com/adg_data_governance_definition/