Data Intensive Research at Raman Research Institute

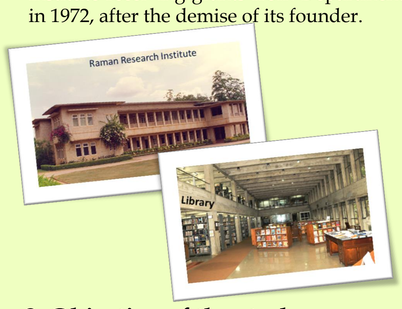
Madhava Rao Meera, Hiremath Vani

Data Intensive Research at Raman Research Institute:
A case study of challenges and perspectives employing questionnaire survey and ETD repository

Meera B M* and Vani Hiremath
Raman Research Institute, C V Raman Avenue, Sadashivanagar Bangalore-560080, India
***meera@rri.res.in**

1. Introduction:

Raman Research Institute (RRI), a pioneering institute of research in physics was started by Noble Laureate Sir C V Raman in 1948 to carry forward his research soon after his retirement from Indian Institute of Science. This self-funded Institute became an autonomous research institute receiving grants from Department of Science and Technology, Government of India, in 1972, after the demise of its founder.



Today, the thrust areas of research at the Institute are Astronomy & Astrophysics; Light & Matter Physics; Soft Condensed Matter and Theoretical Physics. The research activities include work in Chemistry, Liquid Crystals, Physics in Biology, and Signal Processing, Imaging & Instrumentation. RRI, a medium sized research institute has graduate program leading to Doctoral degree in these areas of research. Since 1972, there are 159 theses submitted for the award of Doctoral degree from RRI.

2. Objective of the study:

Douglas Kell (Professor of Bio-analytical sciences) of University of Manchester considers data intensive research as "a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working". Different areas of physics research such as high-energy particle physics or research on nuclear fusion use large data sets. Data intensive research based on huge volumes of data is the current trend and is considered as fourth paradigm in science research.
RRI as an Institute specializing in physics has witnessed data intensive research in the past four decades. This paper is a case study aiming to understand the challenges faced by graduate students and their perspectives in data intensive research at RRI. "Big Data" – being around in the corner, our objective is to find whether research at RRI is heading towards this buzz phenomenon.

3. Research Methodology:

3.1 To understand the perspectives of the past students who have already graduated from RRI, we plan to elicit information by consulting their theses (ETDs) archived in RRI digital repository.

3.2 Our research methodology includes a questionnaire survey of the current graduate students with an aim to understand their perspective regarding the following:

- Data types
- Data collection methods
- Data storage and access
- Data processing challenges
- Legal issues
- Ethical issues/ plagiarism and many more.

4. Data Collection:

This study considers two methods of data collection.

4.1 Employing Electronic Thesis and Dissertations from digital repository of RRI:

Raman Research Institute Library has built a digital repository in 2006 using Dspace as the platform. Essentially, the foremost purpose of creating this repository was to bring global visibility to the published works of Sir C V Raman, the founder of the Institute. Subsequently the repository was also looked up as a tool for showcasing the research publications of the students of Sir C V Raman and later the current research publications of the Institute. As the graduate program progressed at the Institute, more and more print copies of the theses were submitted to the library as a part of the print collection development program.

Theses and dissertations is an important branch of scholarly communication. They are one of the least tapped information resources unlike journal papers or conference papers or for that matter any other primary information resource. Theses and dissertations can be attributed as grey literature which generally does not undergo publishing process. However, it is to be noted here that the content value of scholarship in them cannot be under estimated. Considering the importance of Theses and dissertations, policy decision was taken to include them in RRI digital repository (RRIDR) initiative thus aiming to bring global visibility to them. Theses submitted before 2004 were in print format only. So, they had to be converted into e-format and then uploaded on the RRIDR. Figure -1 is the screen shot of ETD on RRI Digital Repository.

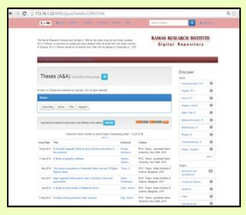


Figure -1 : Screen shot of ETD on RRIDR

Steps involved in building ETD repository:

1. Getting copy right cleared by taking consent from the author of the theses.
2. Identifying metadata
3. Scanning print theses (90 in number) page by page who were not born digital.
4. Making them Optical character Recognition compatible
5. Uploading them on the repository

Post 2004, the entire theses are born digital and the graduate students were requested to submit e-version of the thesis to library. They were directly uploaded soon after defense and copyright clearance. After the award of the degree, thesis is made "open access". As of June 2016, RRI digital repository has 159 theses. For the purpose of data collection of this study, each thesis is consulted to find the following:

- Thesis has data?
- What is the mode of data collection?
- Have they used computer for data analysis?
- Have they employed any software either for data collection or analysis?
- Wherever possible, tried to estimate the volume of data.

4.2 Employing survey questionnaire to elicit perspectives from current graduate students

Survey questionnaire comprising of sixteen questions was administered to 95 current research scholars to get their perspective with respect to research data at RRI.
Data collected from both the methods mentioned above are analyzed in following section.

5. Data Analysis:

5.1 Through ETD of RRIDR

The first method of data collection using ETD repository for the past research from 1972 to 2015 gave us the following results.

5.1.1 Below figure -2 represents total number of theses at RRI Digital Repository, followed by classified representation at the second level giving number of theses submitted by the four different groups. At the third level there is a representation bifurcating "theses with data" and "theses without data". The same information is graphically represented in Figure-3 .

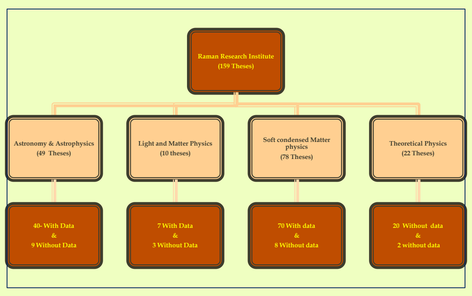


Figure -2 : ETD of RRIDR

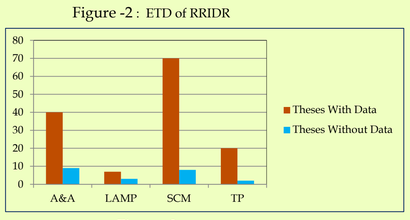


Figure -3

5.1.2 The **mode of data collections** was ascertained by consulting each thesis during the study. It was observed that most of the thesis had employed more than one method of data collection. As a result, there are 211 observed hits for 139 theses. Data obtained is graphically represented in below figure -4.

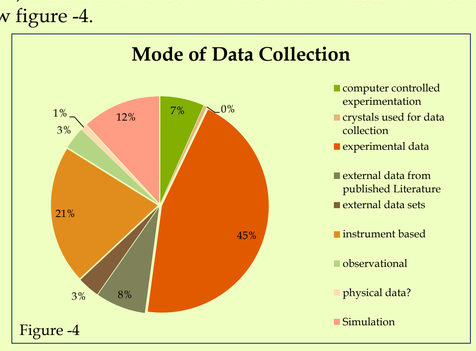


Figure -4

5.1.3 Theses were also consulted to find if the researchers had used computer for **data analysis.** It was apparent that more than 97% have used computer for data analysis followed by 3% of manual method. While doing so, many software's such as DYNALS, CAD, DAQ card, LabView, Mathematica and MBR Software were used by graduate students. They have also used AxioVision software for capturing videos and images and SPIP for image processing.

5.1.4 As regards volume of data, it was almost impossible to find the correct information from thesis. At least 6 of them have mentioned to have used large data sets. There are two theses that make reference to 20,000 hours of observations. Authors of this paper are unable to quantify hours of observations (probably cosmic) in terms of the modern metrics such as bits and bytes. However, there is one thesis which has small data set.

5.2 Questionnaire method

During this survey, we received duly filled in responses from 80 students amounting to 84% response rate. Therefore, our sample size is 80 in this study. Few questions had multiple choice and they were asked to choose all that is relevant.

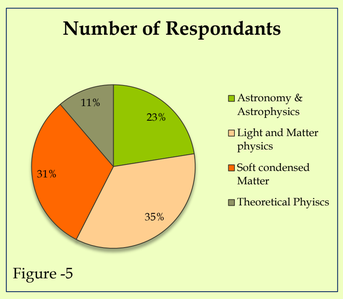


Figure -5

5.2.1 Distribution of respondents with respect to their research group is diagrammatically represented in figure- 5. Out of 80 respondents, presently 51 are in the process of data collection, 26 are yet to begin the process and 3 have unanswered this question.

5.2.2 Second question was to know whether their research is based on using data. Out of 80 respondents, 72 are having data, 6 have no data and 2 have not answered this question. Representation of this data in percentage is given in figure - 6 .

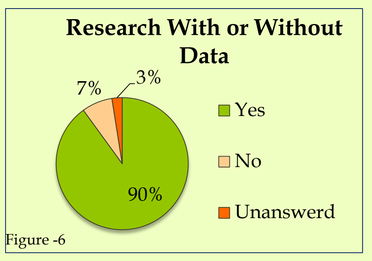


Figure -6

5.2.3 What is the mode of data collection was the next question. We had provided four options and they are a) Instrument based data; b) Observational data; c) Experimental data and d) Simulation data. Respondents were given the freedom to choose all those options that were relevant to them.

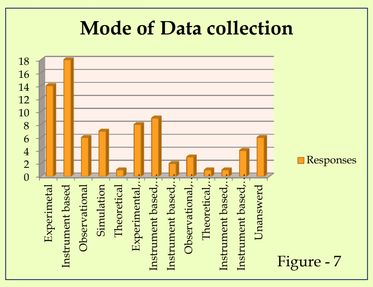


Figure - 7

As a result, we observed different combinations of data collection methods. However, more than 50% are resorting to single method of data collection.
The responses are graphically represented in figure -7 .
From the figure-7 it is very clear that Instrument based research is the most popular category at RRI followed by experimentation. This is an expected result as Astronomy & Astrophysics; Soft Condensed Matter are the two core areas of research at RRI producing maximum number of doctorates.

5.2.4 We wanted to ascertain the data format collected by researchers at RRI, digital being the most popular in recent times. Their response is given in the adjacent pie chart Figure -8 .

Our assumption was to get a response of 100% digital data, and it is surprising to note that 13% are of the view that their data is in non-digital format. Also, 6% have not answered this question.

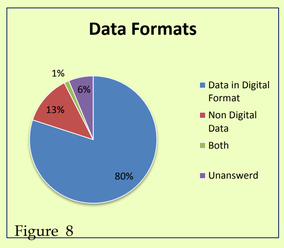


Figure 8

5.2.5 Next question was to find different **types of digital content** that research scholars create, while generating the research data.
We have identified 9 categories of digital content. For each one of them, we have offered more than two options. They were allowed to choose all the relevant options. Figure -9 lists them all along with options. Responses received for these questions are graphically represented in the figure -10

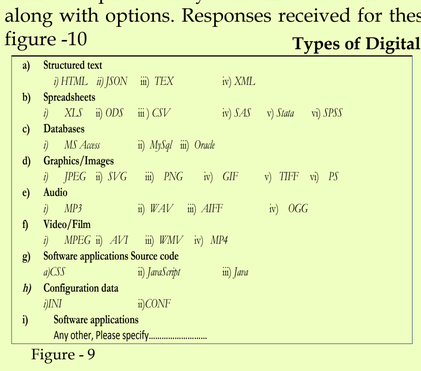


Figure - 9

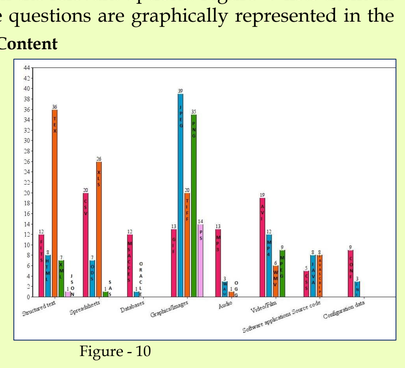


Figure - 10

It is clear from above figure-10 that highest hit is for JPEG followed by PNG in graphics/images category. This seems like JPEG and PNG are the most favored output for image analysis and this is an indication that research happening at RRI is centered on image processing. Another highest peak is observed in structured text category for TeX having 36 hits. It has been an observed norm in all the scientific publications to employ Latex for typesetting system. Our hit results at RRI also confirms to this norm. XLS for spreadsheets is another important digital category, which is generally used for data analysis. This survey endorses this fact by having 26 hits. Rest of the hits are distributed and insignificant for analysis. In addition to these digital content, graduate students at RRI have used many other specific software's such as Mathematica, Labview, Python, Origin Lab, ImageJ, MatLab and many more.

5.2.6 Bigdata and codata are the areas attracting new genera of research in the recent past. So, volume of data is of concern in a study like this while dealing with data and dissertation.

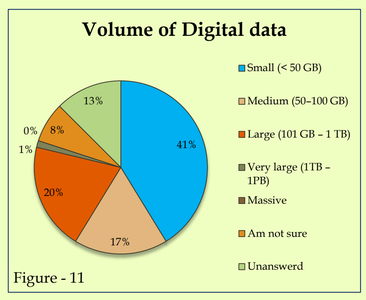


Figure - 11

We had a question to assess the volume of data that will be collected by our scholars in the process of doctoral research. To our surprise, as per figure -11, maximum hits is for small data disproving our hypotheses. As per this, we are forced to come to a conclusion that data intensive research is not happening at RRI. It was hypothesized that research at RRI would be heading towards big data for the very reason that astronomy & astrophysics and soft condensed matter being two core areas of research, which generally tend to be data intensive.

5.2.7 Data storage facility is a major concern while handling big data. 85 % of the respondents have expressed their satisfaction with respect to the storage facility available at RRI. They use multiple devices to store research data. Their response with respect to data storage is represented graphically in the adjacent figure -12. Storing research data on personal/workplace computer is the obvious choice which is preferred by 46% of the respondents followed by External hard drive such as USB/CD/DVDs. The other three choices are marginally used by scholars.

Figure - 12

5.2.8 Skills such as a) programming language b) knowledge about different software's and c) statistical tools are prerequisites for research while handling data. They play a major role both during data collection and data processing in all disciplines of research. Research Scholars were asked to choose all the relevant options. Therefore, we have 8 options and they are represented below in the pie chart Figure -13.

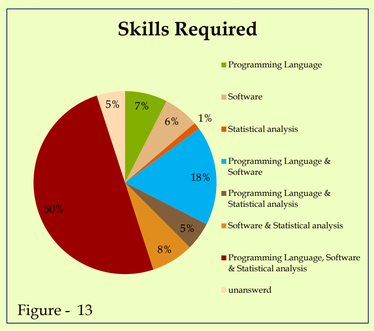


Figure - 13

50% of the respondents have expressed their opinion that all the three skills are important. Programming language and software are considered to be important skills by 18% of the people. Hits for the other combinations of skills or independent skills are almost identical and hence has less scope for analysis.
70% of the respondents are knowledgeable about the skills required. 22% of them have expressed their desire for training in programming languages.

5.2.9 Data processing is an important aspect of research. It could be done either manually or the process can be automated. A question to find the choice at RRI fetched us the following result represented through pie chart in figure-14.

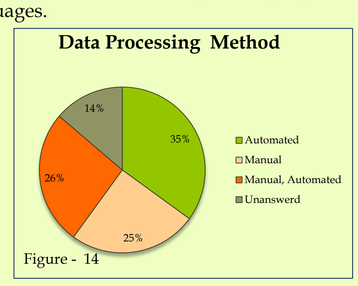


Figure - 14

For the purpose of automated data processing graduate students have used **variety of software's**. The number wise distribution of software's used for data processing are represented graphically in figure -15 .

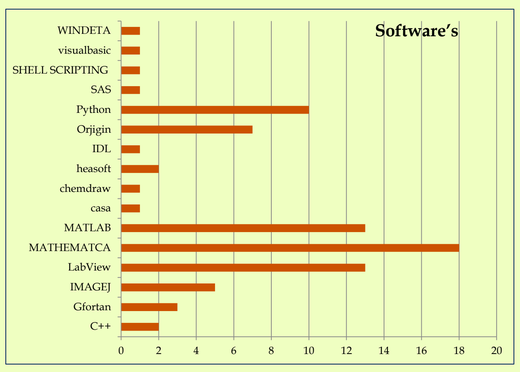


Figure – 15

5.2.10 Archiving research data for posterity is widely accepted especially in disciplines like Astronomy & Astrophysics and nuclear physics. During this survey when we sought their opinion, 47% percent were not in favor of archiving, 40% supported archiving and the rest 13% left it unanswered. All those who are in favor, have voiced their choice as "self archiving".

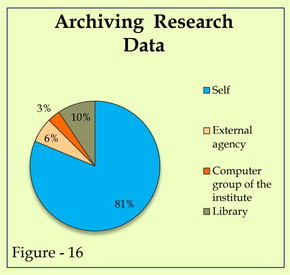


Figure - 16

5.2.11 Using **External data** from repositories and published data sources are well accepted phenomenon in Science research. Majority of the graduate students (77%) in this survey are not using external data of any kind. The balance 23% (16 Respondents) use external data. 14 out of 16 are from A& A group of research and rest 2 are from SCM group.

5.2.12 "**Research data loss**" is a cause of concern. Data can be lost inadvertently, and measures need to be taken to handle such situations. There are 69% of researchers who have not experienced data loss, indicating conducive research atmosphere at RRI. There are 60 % of them who have protected data by using password protected, limited access computers. The other options such as data encryption, working in limited access secured data rooms and working on machines without internet connectivity are the measures taken by very few respondents which does not gives scope for analysis.

5.2.13 Next question was to get the opinion of researchers regarding **sharing of research data**. There are 52% of the respondents who have said 'yes 'to share the research data and they share data with their research guides, Lab/ Group members and collaborators. Data sharing pattern is represented in the adjacent figure -17 .

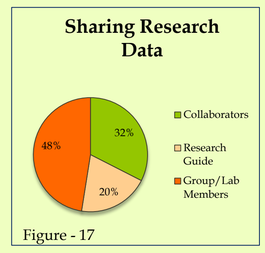


Figure - 17

5.1.14 While sharing / using external data, **legal issues** are of major concern. We wanted to know the perception of our graduate students about this aspect which is very important. To our surprise, majority of them are unaware of the legal clauses that are associated in data sharing process. Out of 80 respondents only 14% have heard about legal aspects, without having any clarity. So we suggest an awareness program about legal aspects of sharing research data.

5.1.15 Moral responsibilities of researchers is an important attribute, be it science or social science. This has been of great concern during the last 3-4 decades where data surge has grown tremendously with the advent and usage of internet and many other ICT tools. Since the objective of our present study is to understand the perspectives of researchers regarding data and dissertation, we felt the necessity to learn their understanding of moral responsibilities such as **a) Ethical Issues, b) Plagiarism and c) Data citation attribution**. Our question as to know whether they are aware of their moral responsibility fetched us the following result represented in Figure - 18

In the recent past, RRI had a directive from Jawaharlal Nehru University, New Delhi, India to which RRI is affiliated for the award of doctoral degree. As per the directive, it has become mandatory for all the thesis to pass through plagiarism check before they are submitted to the award of degree. As a result majority of our graduate students (80%) are aware of plagiarism. However, there are 13% of respondents who are not aware of plagiarism and balance 7% who have not answered the question. We presume 13% are fresh graduate students.

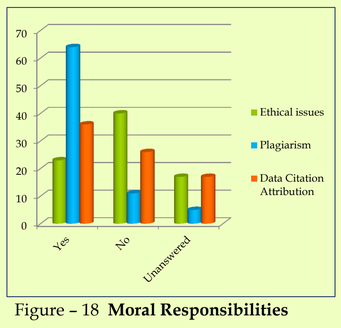


Figure – 18 **Moral Responsibilities**

Data citation attribution is a procedure of acknowledging the use of any external data. This is quite similar to citing referred information while one is in the publishing process. Data citation attribution is fairly a new concept in science research and to our surprise there are 45% of respondents who are knowledgeable about this.
Half of the respondents are not aware of ethical issues in data intensive research. Ethical issues, probably would have been addressed already inherently by the graduate students and if not, by their experienced mentors without making an issue about it. So the end result is justified.

5.1.16 Our respondents were questioned about the kind of support options that they expect from the institute while handling research data. They were asked to choose all the relevant options, so that we get a clear idea about their requirements. Their response is given in figure -19. The majority of graduate students have felt the need for better technical infrastructure followed by training courses and data processing facilities. The fourth position is taken by data management followed by ethical issues, legal advice and creating data.

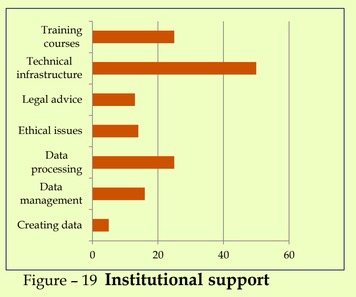


Figure – 19 **Institutional support**

Finally, we wanted to know, if the current graduate students wish to share any additional information about data and research. There has been a request for an online data storage system for the entire institute, quite similar to cloud facility. There are few respondents who have suggested data acquisition software and training associated with that. However not many research scholar have taken the trouble of expressing their wish , which otherwise would have thrown some light for further analysis.

Inferences

1. Our hypothesis was that research at RRI is increasingly becoming data intense. But, the present study disproves our hypothesis and it indicates that research centers around "**Small Data**" at RRI.
2. The most favored data collection methods are **instrument based and experimental methods,** thus proving our hypothesis.
3. It was also hypothesized that the data intensive research critically hinges more upon ease of access to data which in turn depends on facilities like **storage capacity, download speed, data processing and computational facility** available, and not so much on legal or ethical issues, as the later would have been addressed already by the graduate students and if not, by their experienced mentors.
   Our data positively respond to this, as our respondents are satisfied with all the **infrastructural facility at the institute**. They are also knowledgeable about variety of skills required for the data collection and analysis.
   As hypothesized, the ethical issues are not of much concern amongst researchers as they are well built into the system.
4. Our next hypothesis was about tools and techniques and their major role in data related research. This hypothesis is proved as our graduate students have responded that they use many **software's, programming languages and statistical tools and techniques** for data analysis.
5. As regards **data sharing and collaboration network** in basic science research, our data indicates 52% of respondents being favorable. So logistically our hypothesis is proved.

Conclusion

These inferences are drawn based on the analysis of the responses received in the questionnaire survey and the information captured from the ETD repository of RRI. Since our first hypothesis is disproved, we are forced to conclude that research data at RRI is not heading towards – "**Big Data".**

A study like this will help in understanding the challenges faced by graduate students in a research institute of medium size. Majority of the graduate students have expressed their desire for a better technical infrastructure and they wish for training courses on different aspects of data related research. Legal aspects, ethical issues, Plagiarism and data citation attributions are some of the social issues associated with data and research. Our observations in this study indicate that researchers are not well informed about these issues. However trivial they may be, research scholars needs to be informed as they come a long way in their scientific pursuit.