Annals of Library and Information Studies
Vol. 65, March 2018, pp. 43-49

Data-intensive research in physics: challenges and perspectives

Meera B M[a] and Vani Hiremath[b]

[a]Librarian, Raman Research Institute, Sadashivanagar P O, C V Raman Avenue, Bangalore – 560080,
Email: meera@rri.res.in
[b]Library Assistant, Raman Research Institute, Sadashivanagar P O, C V Raman Avenue, Bangalore – 560080,
Email: vanih@rri.res.in

Received: 2 March 2017; revised: 7 March 2018; accepted: 20 March 2018

Raman Research Institute (RRI) in recent years has been engaged in data intensive research. The paper aims to understand the challenges faced by graduate students and their perspectives in data-intensive research at RRI regarding data types; collection methods; storage and access; data processing; legal and ethical issues. Questionnaire survey method and information extraction from ETD repository of RRI were used for data collection. Graduate students have expressed their desire for a better technical infrastructure, and need for training courses on different aspects of data related research and acquisition of data processing software. Observations in this study indicate that researchers are not well informed about some of the social issues associated with data and research such as legal aspects, ethical issues, plagiarism, data citation attributions, etc,. Study recommends steps to empower research students to handle the challenges in data intensive research.

Keywords: Research data; Big data; Data type; Data storage; Data processing

Introduction

Data intensive research is considered 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. Data-intensive research based on large data is the current trend and is considered as the fourth paradigm in science research."[1]

The Raman Research Institute (RRI), Bengaluru, India, founded by Indian physicist and Nobel laureate Sir C. V. Raman, is a pioneering institute pursuing research in basic sciences. Sir C. V. Raman setup RRI to carry forward his research soon after his retirement from the Indian Institute of Science in 1948. This self-funded Institute later became an autonomous research institute in 1972 and started receiving grants from Department of Science and Technology, Government of India. Today, the thrust areas of research at the Institute are: 1) Astronomy & Astrophysics, 2) Light & Matter Physics, 3) Soft Condensed Matter and 4) Theoretical Physics. RRI, a medium sized research institute has a graduate program leading to Doctoral degree in these areas of basic science. Currently, there are around 40 faculty and 95 graduate students pursuing research at RRI.

RRI presently has nearly seven decades of research history in basic sciences. Around 4000 research papers have been published till recently in prestigious peer-reviewed journals. Since 1972, there are 159 theses submitted for the award of Doctoral degree from RRI. Since RRI is not a deemed university, the doctoral degree is awarded by Jawaharlal Nehru University, New Delhi. Different areas of physics research such as high-energy particle physics or research on nuclear fusion, astronomy and astrophysics use large data sets. The aim of this study is to figure out the role of 'data' in current research activities at RRI.

Review of literature

Data-intensive research, big data, open data, and co-data have become popular in the recent past in all domains of research, be it physical science, bio-science, humanities or social science. Data generated in different walks of life is subject to rigorous computational analysis, and they are expected to churn out fruitful results in handling complex life situations. It could be medicine, law, education or any other subject where data analytics is involved. The EDUCAUSE Center for Analysis and Research

(ECAR) report in 2005 has dealt with the topic "Rise of data-intensive research" manifesting a vision of how data-intensive research will transform both science itself and the social organization of how that science is conducted[2]. Lynch, in his paper, "Big data: how do your data grow?" has addressed various issues associated with data preservation and management, considering the huge amount of investment that goes into international projects such as CERN or the Large Synoptic Survey Telescope etc[3].

Howe and others have felt the need for data curation and preservation of raw biological data. They suggest how authors, publishers, and curators should come together to make this happen[4]. Boyd and Kate have discussed the emerging trends in big data studied by computer scientists, physicists, and economists and have tried to study the cultural, technological and scholarly phenomenon in the realm of big data[5]. McDonald and Leveille have identified the issues associated with retention and disposition specifications within the context of big data and open data initiatives based on the analysis of business processes and workflow[6]. Childs, Sue and others have highlighted the role of three key issues namely methodological issues; ethical and practical issues which should be considered while making research data open and accessible for sharing and reuse. Additionally, they have identified new roles and opportunities for records managers in the open data and RDM contexts[7].

Jin, Xiaolong and others have defined and discussed the features and value of big data in the light of Internet, Internet of things (IOT) and cloud computing. They have also identified the challenges namely data complexity, computational complexity, and system complexity and possible solution to address them[8]. Curdt and Hoffmeister have addressed issues concerning design and implementation of RDM system in interdisciplinary studies with particular reference to soil-vegetation-atmosphere data[9].

Anagnostopoulos and others have classified some of the critical challenges of handling big data, and they have also recommended solutions and also highlighting directions for cross-disciplinary research[10]. Muller and others have examined big data analytics for information systems research and in the process; they have also come up with a set of guidelines for conducting big data analytics in information science. They have also introduced some methodologies and tools that are not familiar to information science professionals[11].

In light of this, we attempt to address the issues associated with data-intensive research faced by the graduate students in the field of physics, where now Big Data has become relevant. The modes of data collections, data storage and processing in a medium-sized research institute with limited facilities are addressed in this study. The resources available to tackle legal and technical issues are also of concern in the context of a physics research institute in India.

Objectives of the study

- To understand the challenges faced by graduate students and their perspectives in data intensive research;
- To investigate if 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;
- To assess the knowledge and awareness of graduate students regarding soft skills such as legal or ethical issues which are necessary while conducting data intensive research

Methodology

This study has employed two types of research methodology.

The first method is aimed to understand the data collected by past students who have already graduated from RRI. For this purpose, each of the 159 Ph.D. degree Electronic Theses and Dissertations (ETDs) since 1972 archived in RRI digital repository was consulted to find if the theses had data and if yes, the type of data, how it was collected, the extent of computers used for data analysis, estimation of volume of data, etc.

In the second method, questionnaires containing 16 questions were administered to 95 current research scholars with an aim to understand their perspectives regarding the following:

- Data types
- Data collection methods
- Data storage and access

- Data processing challenges
- Legal issues
- Ethical issues/ plagiarism and much more.

Information collected through both the methods has been analyzed in the following sections.

Analysis

Method 1: Through ETD of RRI digital repository

Table 1 gives information on theses in RRI Digital Repository with and without data under various groups and departments of the Institute. It is seen that 86% of the theses had data.

It is observed that most of the thesis had employed more than one method of data collection. A total of nine data collection methods were seen in the 137 theses that had data (Fig. 1). Experimental data collection is seen as the prominent mode of data collection with 45% of the researchers using this method to collect data. This is followed by instrument based data collection. The other modes included simulation, external data from published literature, computer controlled experimentation, observational data, external data sets, physical data, etc.

It was also seen that 97% of the thesis contain computer based data analysis. Software such as DYNALS, CAD, DAQ card, LabView, Mathematica, and MBR Software were used in the theses. They have also used AxioVision software for capturing videos and images and SPIP for image processing.

As regards volume of data, it was not possible to find the exact information from the theses. At least 6 of them have mentioned having used large data sets. There are two theses that make reference to 20,000 hours of observations. We were unable to quantify hours of observations (probably cosmic) in bits and bytes.

Method 2: Questionnaire Method

Out of the 95 questionnaires administered, 80 duly filled-in questionnaires were received amounting to a response rate of 84%. From the responses it could be seen that 90% of the respondents were using data, 7% do not use data no data, and 3% have not answered this question.

Modes of data collection

Out of the four modes of data collection provided in the questionnaires viz., instrument based data,

Table 1—Theses in RRI Digital Repository

Groups/Departments	No. of theses with data	No. of theses without data	Total no. of theses
Astronomy & Astrophysics	40	9	49
Light and Matter Physics	7	3	10
Soft Condensed Matter	70	8	78
Theoretical Physics	20	2	22
Total	137	22	159

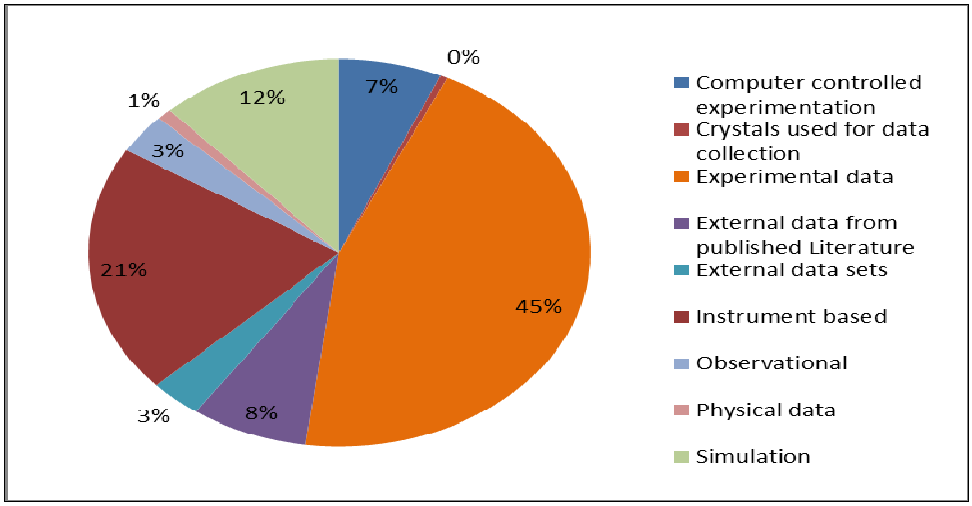

Fig. 1—Mode of data collection

observational data, experimental data; and simulation data, it is found that instrument based research is the most popular category opted by 17 researchers, followed by experimental method (13 numbers), observational methods (5 numbers), and simulation method (6 numbers). The fifth option – theoretical method was provided by one respondent. Different combinations of the mode of data collection were chosen by the rest 38 graduate students.

Expectedly, 80% of the respondents collect data in digital form. As for types of digital content, data collection in the graphics/image category in JPEG file format followed by PNG were the most common. There is considerable image processing research going on at RRI and that perhaps explains the prevalence of these two file formats. Data in structured text category for based on LaTex program is also is used considerably followed by XLS for spreadsheets.

Fig. 2 gives the responses with regard to the volume of data.

Clearly, big data research is not yet common in RRI although it was hypothesized that astronomy & astrophysics and soft condensed matter being two core areas of research in RRI, and which tend to be data intensive, might see the use of big data. However, 41% of the responses indicate less than 50GB data, followed by 20% that have large data sets.

Responses about data storage facility is given in Fig. 3.

It is seen that 85% of the respondents have expressed their satisfaction about the storage facility available at RRI. They use multiple devices to store research data. Storing research data on personal/workplace computer is the choice preferred by 46% of the respondents followed by the external devices such as USBs and optical disks. The other three options, namely – directly on the machine or system, a central server of the Institute and external data centre are marginally used by scholars.

Data processing is an important aspect of research. It could be done manually, or the process can be

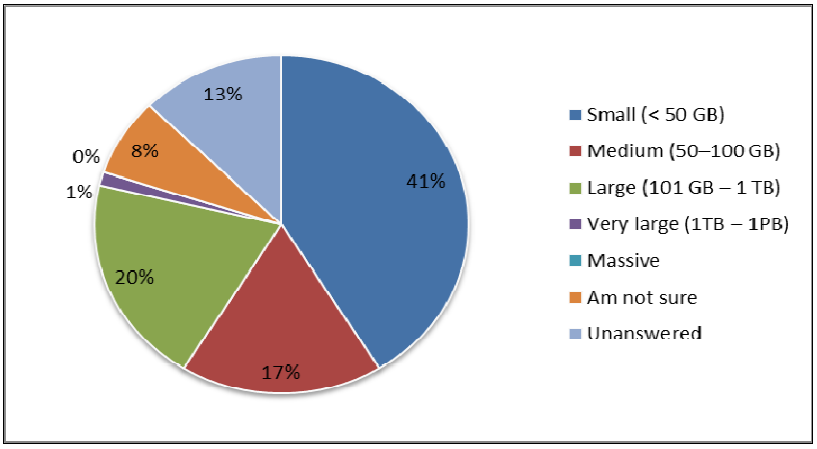

Fig. 2—Volume of digital data

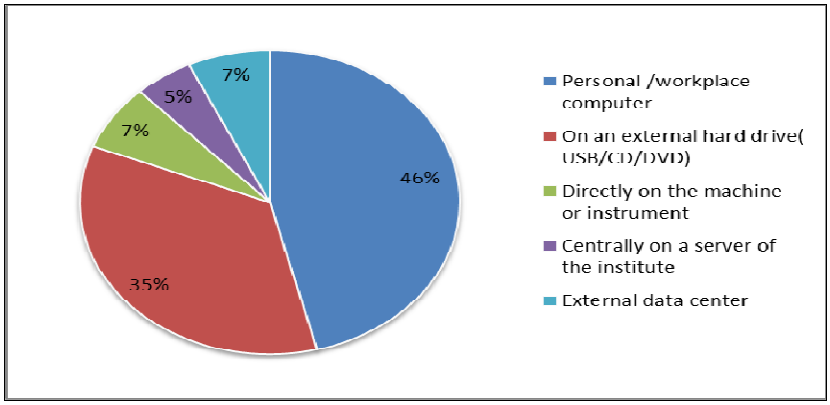

Fig. 3—Data storage

automated. A question to find the choice at RRI fetched result represented in Fig. 4.

Automated data processing is the choice of 35% respondents, followed by 26% respondents who are using both automated and manual methods. However, there are about 25% whose choice is manual and 14% who have not answered the question.

For the purpose of automated data processing, graduate students have used a variety of software. The distribution of software used for data processing is represented graphically in Fig. 5.

The software Mathematica is mostly used followed by MATLAB and LabView.

Archiving research data for posterity is common in the disciplines of astronomy, astrophysics and nuclear physics. From Fig. 6 it is seen that 81% carry out self-archiving and a very small percentage take the help of external agency, IT facility and library of the institute.

Research data loss is a cause for grave concern. Data can be lost inadvertently, and measures need to be taken to handle such situations. About seventy percent of researchers have not experienced data loss at RRI. Sixty percent of the respondents have protected data by using passwords. The other options such as data encryption, working in restricted access secured data rooms and working on machines without Internet connectivity are the measures taken by a few respondents.

Opinion of researchers regarding sharing of research data was also obtained and it was found that only 52% of the respondents who share data. Legal issues are involved when sharing data or using external data. It was noted that majority of the respondents were unaware of the legal clauses that are associated with the data sharing process. Out of 80 respondents, only 14% have heard about legal aspects.

Since the objective of the present study is to understand the perspectives of researchers regarding data and dissertation, it was also felt necessary to find out their understanding of plagiarism and data citation attribution. Eighty percent of the RRI research students are aware about plagiarism. Data citation attribution is a procedure of acknowledging the use of any external data. It is quite similar to citing referred information while one is publishing a research paper. Data citation attribution is fairly a new concept in science research, and 45% of respondents who are knowledgeable about this.

The 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 it is possible to get a clear idea about their requirements. The majority of graduate students specified the need for better technical infrastructure followed by need for training courses and data processing facilities. The fourth position is taken by data management followed by ethical issues, legal advice and creating data.

Finally, the study aimed to find if the current graduate students wish to share any additional information about data and research. There has been a

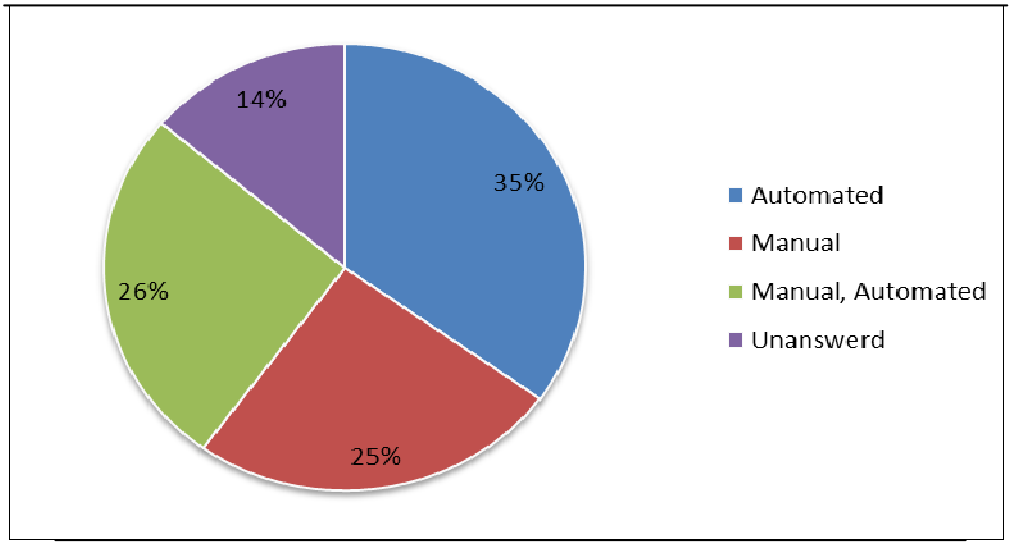

Fig. 4—Data processing

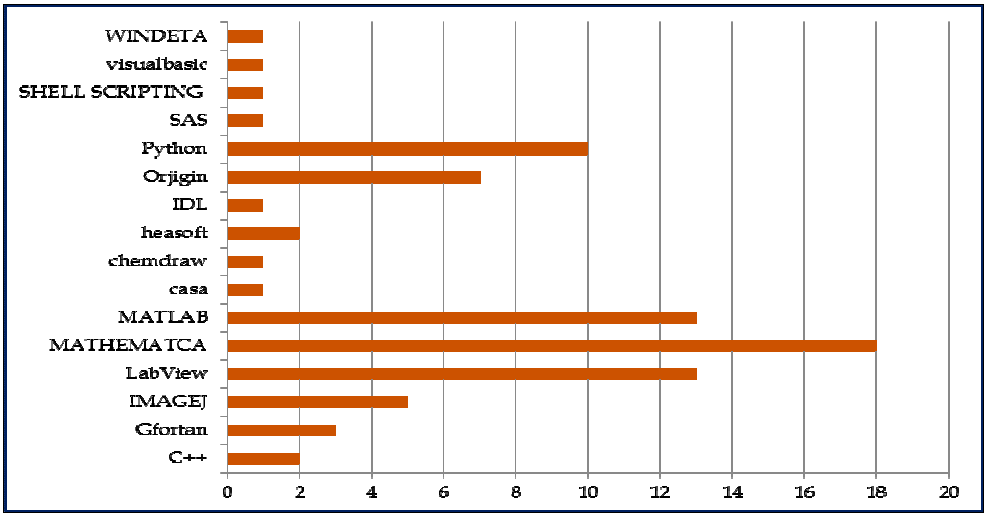

Fig. 5—Data processing software

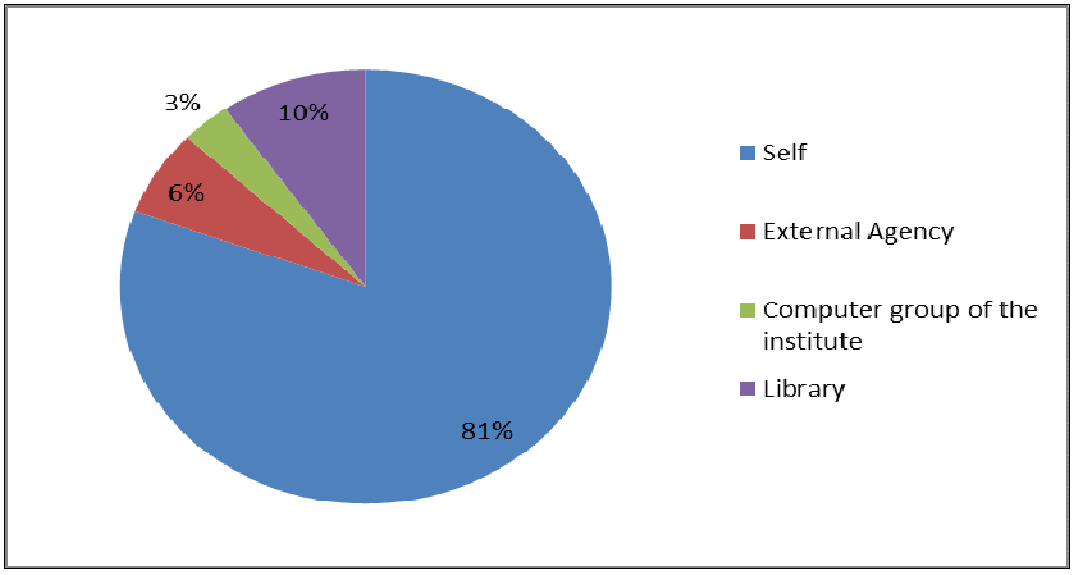

Fig. 6—Archiving research data

request for an online data storage system for the entire institute, quite similar to cloud facility. A few respondents have suggested data acquisition software and training associated with that.

Conclusions

From the study it can be concluded that, although RRI is involved in data intensive research, there is very little research on big data. With respect to data, there seem to be a need for better technical infrastructure, training courses on different aspects of data related research and mostly importantly, more awareness of the legal and ethical aspects in connection with data handling.

References

1. Hey A J G, Tansley S and Tolle K M, The fourth paradigm: data-intensive scientific discovery (Microsoft Research, Redmond, WA) 2009

2. The rise of data-intensive research – Educause. Available at https://net.educause.edu/ir/library/pdf/ers0605/rs/ers06054.pdf. (Accessed 20 March 2017).

3. Lynch C A, Big data: How do your data grow? *Nature*, 455 (7209) (2008) 28-29.

4. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill D P, Kania R, Schaeffer M, et. al. Big data: the future of biocuration, *Nature*, 455 (7209) (2008) 47-50.

5. Boyd D and Crawford K, Critical questions for big data, *Information, Communication and Society*, 15 (2012) 662-679.

6. McDonald J, and Leveille V, Whither the retention schedule in the era of big data and open data? *Records Management Journal*, 24 (2) (2014) 99-121.

7. Childs S, McLeod J, Lomas E, and Cook G, Opening research data: issues and opportunities, *Records Management Journal*, 24 (2) (2014) 142-162.

8. Jin X, Wah B W, Cheng X, and Wang Y, Significance and challenges of big data research, *Big Data Research,* 2 (2) (2015) 59-64.

9. Curdt C, and Hoffmeister D, Research data management services for a multidisciplinary collaborative research project: Design and implementation of theTR32DB project database, *Program*, 49 (4) (2015) 494-512.

10. Anagnostopoulos I, Zeadally S and Exposito E, Handling big data: research challenges and future directions, *The Journal of Supercomputing*, 72 (4) (2016) 1494-15160.

11. Muller O, Junglas I, Brocke Jan V, and Debortoli S, Utilizing big data analytics for information systems research: challenges, promises and guidelines, *European Journal of Information Systems*, 25(4) (2016) 289-302.