Review

Supporting e-Science: Scientific Research Data Curation

Radovan Vrana
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb,
Ivana Lučića 3, Zagreb, Croatia
rvrana@ffzg.hr

Summary

One of the outcomes of the continuous development of science is creation of new methods of scientific research which, as a result, generate different types of research output including research data. While a significant attention is given to preservation of journal articles, books, and papers published in conference proceeding, less attention is given to the preservation of research data. To enable use of data accumulated in previous scientific research projects in a new scientific research, research data should be preserved. The activity of research data preservation is called data curation. Data curation has become necessary if science wants to avoid data loss. Unfortunately, science itself cannot take care of research data easily; it needs help from professionals like librarians and archivists to preserve research data in order to enable their re-use future scientific research. Although there are already some good solutions to this problem such as storing research data in digital repositories, no final decision has been made about who will take the responsibility for this kind of activity in the long run.

Keywords: scientific communication, e-science, data curation

Introduction

One of the outcomes of the continuous development of science is a creation of new methods of scientific research which, as a result, generate different types of research output including research data. While usual scientific output in form of journals articles, books, and conference papers is preserved and made available in form of local digital collections in libraries owned by academic institutions or in remote full text databases for access to which academic institutions are paying licenses, research data is rarely preserved. The research data have a special value, since they can be included into new scientific research projects and lead to new scientific discoveries. To enable such use of data accumulated in previous scientific research projects, research data should be preserved. As this

problem of preservation of research data attracted more attention over the years, scientific community in cooperation with libraries and archives started to consider development of a support in form of procedures, policies, guidelines and standards for long term preservation of research data in various formats. In addition to these written documents, a new activity of long term preservation of research data was initiated and it became known as data curation. The purpose of this activity is to facilitate re-use of research data created as output of a previously completed scientific research in a new scientific research. Inclusion of research data in the new research has become possible because of the changing role of data in the scientific world. Research datasets have ceased to be merely the output of the research endeavour, and they have become a new input to new hypotheses which enable new scientific insights and drive innovation (National Science Foundation, 2007). Consequently, it has become necessary to enable systematic capturing and preservation of the scholarly output with a special attention given to the scientific research datasets in order to minimise the risk of data loss in an ever-changing environment "(...) where data flows and technologies are changing constantly." (Angevaare, 2009, p. 4). The risk of data loss is becoming greater as our reliance on digital information resources grows as well as our negligence about the future of the research output in digital form: "By the time knowledge in digital form makes its way to a safe and sustainable repository, it may be unreadable, corrupted, erased, or otherwise impossible to recover and use." (Ogburn, 2010, p. 242). Ogburn (2010, p. 242) offered reasons why research data may be endangered: "(…) due to their sheer size, computational elements, reliance on and integration with software, associated visualizations, few or competing standards, distributed ownership, dispersed storage, inaccessibility, lack of documented provenance, complex and dynamic nature, and the concomitant need for a specialized knowledge base—and experience—to handle data." Data curation has also become necessary because the scientific endeavour has become too expensive to let research data be destroyed or forgotten. Some research data are unique and cannot be recreated which make them primary concern for their long term preservation. Science itself cannot take care of research data easily; it needs help from professionals like librarians and archivists to preserve research data in order to enable their re-use future scientific research. This paper will give an overview of the problem of data curation and emphasize its significance to the modern science.

E-science, research data and data curation

Nowadays, when we speak about the modern science, we usually refer to ICT supported networked science that is global, data intensive and more collaborative than ever. We call such science - e-science. According to Lord and MacDonald (2003, p. 5) "The term e-Science – or more inclusively e-Research - has been used recently to describe the research culture and opportunities enabled by

these developments, and the collaborations of people and of shared resources that are needed to resolve new research challenges, whether in the sciences, social sciences or humanities. E-science enables a new order of collaborative, more inter-disciplinary research, based on shared research expertise, instruments and computing resources, and crucially increasing access to collections of primary research data and information - the knowledge base of research.". E-science is characterized by a creation of great quantities of data "(…) generated from sensors, satellites, high-performance computer simulations, high-throughput devices, scientific images and so on (…)", and they will "(…) soon dwarf all of the scientific data collected in the whole history of scientific exploration." (Hey and Trefethen, 2003, p. 4). To facilitate global collaboration, e-science needs an infrastructure that will make possible "(…) sharing of computing resources, data resources and experimental facilities in a much more routine and secure fashion than is possible at present." (Hey and Trefethen, 2003, p. 1). Digital technology should provide such infrastructure and secure long term preservation of "(…) data generated today so it can survive the changes of technology and can be accessed in the future." (Hockx-Yu, 2006, p. 234). Taking care of the research data in the long term will be a very difficult task as it requires complete understanding of versatile and complex digital objects, procedures which create these data, procedures to decide what data to keep, hardware and software necessary to complete data care, trained staff that will operate digital repositories of research output etc.

Different scientific fields have different types of research output. When speaking about the long term preservation in science (in general), one usually refers to the preservation of printed scientific output such as journal articles, books, and conference papers published in proceeding but also to research data. The term data is here used "(…) to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment." (National Science Foundation, 2005, p. 9). For Beagrie, Chruszcz, and Lavoie (2008, p. 18) "Research data is an essential input to scholarly endeavour, whether that endeavour is focused on extending the frontiers of knowledge, or understanding the discoveries of the past. (…)". To make use of such research data, e-science needs a well organized middleware infrastructure which includes a possibility of making available data created by the scientific research on demand.

Collecting, organizing and processing research data and facilitating their re-use are activities related to data curation. The term data curation is rather new. Its life began at the beginning of the 1990s when a need for more specific preservation of different types of material or objects had arisen. For instance, museums and libraries need care for physical artefacts while researchers in natural sciences need care for databases comprising data of the human genome.

(Beagrie, 2006, p. 5). For Shreeves and Cragin (2008, p. 89) "Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation and organization of these data for access and use over time". Van Horik (2008, p. 132) explains that there is a subtle difference between data curation and long term digital preservation: "Curation not only implies the preservation and maintenance of a collection or database, but also relates to the creation of added value and knowledge." British Digital curation centre offers similar view of digital curation: "Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle." (Digital curation centre, 2010). According to Lord and Macdonald (2003, p. 12), data curation is: "The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.". All these views on data curation suggest existence of an understanding about major issues surrounding data curation, which will lead to finding better solutions for execution of this important activity.

Why is data curation needed?

Data curation is needed for several reasons. Most new researches refer to the findings of the previous researches i.e. data from the previous researches. "Data are evidence supporting research and scholarship; better research is based on verifiable data, which may in turn lead to new knowledge." (Digital curation centre, 2010). By using the output of the previous researches, researchers continue the work of other colleagues who made discoveries before them. That is the main principle how science works. With the growth in volume, complexity, and heterogeneity of digital information, the requirement for active management becomes more challenging and more critical to a wider range of organisations (Van Horik, 2008, p. 133). In addition to these changes, we are witnessing introduction of new electronic devices of different types which provide support to scientists during their research, and produce new research data. As a result, it has become very difficult to collect, organize and process huge amounts of very different types of data and to find purpose for their preservation and re-use. The volume of data created by the scientific research is constantly increasing at a staggering rate, and it has already become huge, so, to secure successful re-use of research data in future, it has become essential to develop technical, organizational, financial and other means to deal with large quantities of research data. In the next decade we will see new experimental facilities coming online that will generate data sets ranging in size from 100's of Terabytes to 10's of Petabytes per year." (Hey and Trefethen, 2003, p. 3). For instance, the Large

Hadron Collider (LHC) at CERN (Geneva) generates roughly 15 petabytes of data annually from 2007 (Van Horik, 2008, p. 133). These volumes of research data are the reason why we need a well planned activity of preservation which would facilitate their future use. Properly curated research data in digital format can be then readily integrated into a new research and learning workflows now and in the future. The research community will benefit from research data curation in four main ways: "1. Improving access. Digital curation procedures allow continuing access to data and improve the speed of access to reliable data and the range of data that can be accessed. 2. Improving data quality. Digital curation procedures assist in improving data quality, improving the trustworthiness of data, and ensuring that data are valid as formal record (such as the use as legal evidence). 3. Encouraging data sharing and reuse. Digital curation procedures encourage and assist data sharing and use by applying common standards and by allowing data to be fully exploited through time (thus maximizing investment) by providing information about the context and provenance of the data. 4. Protecting data. Digital curation procedures preserve data and protect them against loss and obsolence.". (Harvey, 2010, p. 12). Joint Information Systems Committee (JISC) (E-Science Data Curation, 2004) enumerated "(…) seven reasons to keep data: re-use of data for new research, including collection-based research to generate new science; retention of unique observational data which is impossible to re-create; retention of expensively generated data which is cheaper to maintain than to re-generate; enhancing existing data available for research projects; for compliance with legal requirements; to validate published research results and for use in teaching".

Not all data created by scientific research will have long term value. Those data that will have long term value must meet several conditions (Van Horik, 2008, p. 139): "1. Digital research data must be findable by means of a catalogue on the internet. This makes appropriate documentation of the research data relevant. 2. Digital research data must be accessible, provided that privacy rules and intellectual property conditions are taken into consideration. The ultimate goal is to realise open access to the research data, 3. Digital research data must be available in a useable data format, enabling secondary analysis in the future. Therefore the research data must be available in a format that can be processed by common available hardware and software, now and in the future. 4. Digital research data must be reliable, that is, the research data is authentic and not changed in the course of time. 5. Digital research data must be referable in a durable manner. This implies that the research data is provided with persistent identifiers and stored in a in a so-called trusted digital repository.".

During his research related to research data curation, Harvey (2010, p. 56), concluded that "(…) digital curation aims to produce and manage data in ways that ensure they retain three characteristics: longevity, integrity and accessibility. Longevity refers to the availability of the data for as long as their current and future users require them; integrity refers to the authenticity of data – that

they have not been manipulated, forged or substituted; authenticity requires that we can locate and use the data in the future in a way that is acceptable to their designated community". These and other authors contributed to the growing body of knowledge about the development of infrastructure for modern networked science and its particular parts which are of great importance to all members of the global scientific community, but also to the society which benefits directly from the scientific endeavour. As science continues to develop, issues like research data curation will attract even more attention than today. The next part of the paper focuses on taking the responsibility for research data curation.

Whose responsibility is it?

Keeping the research data safe is a very responsible job. Scientific community is currently seeking individuals (less likely to take this role), government institutions (more likely to take this role) and commercial enterprises (also more likely to take this role) which are willing to take the role of data keepers.

Scientists are the most responsible for the created research data as they are responsible for carrying out of research projects which produce the research output including data from different measurements, observations and experiments. In the conclusion of her report on dealing with research data, Lyon puts focus on institutional and human aspects of data curation activities. Concerning scientists, the human aspect of data curation, Lyon (2007, pp. 59-60) claims that "(...) many researchers appear to be unaware of the range of issues associated with data management best practice and there is a growing requirement for coordinated advocacy, training and skills programmes to equip the research community with the appropriate competencies to foster the Science Commons envisaged for the future.". Naturally, there are differences among researchers. Jones (2003, p. 3) distinguished three different categories of researchers regarding data curation: "(…) those who don't trust the ability of digital repositories to take care of their material; those who are unaware that such a possibility exists; and those who would love to be able to hand over their materials but no obvious repository yet exists for them to do so.". Generally, to help research data capturing, preservation and exchange, scientists should do the following: apply open-source software and open standards to encourage interoperability among different software and hardware platforms; create metadata and annotations so that digital objects can be reused; link related research materials and make sure the links are persistent; use persistent identifiers; be consistent about citation formats; decide which digital objects need to be curated over longer term, keep data storage devices current; validate and authenticate migrated data. (Harvey, 2010, p. 58). There are many incentives for sharing data with other scientists: publicising the results of their research which occasionally includes data, requirements made by publishers that data underlying an article is to be

made available on request to other researchers, agreements that require data sharing with other research projects in the same area, adherence to open access principles etc. (Ruusalepp, 2008, p. 5).

Because of their position in the research community, academic and research libraries are frequently mentioned as institution where results of the scientific endeavour are kept. Shearer and Argaez (2010, p. 3), suggest that libraries are well positioned to support research data stewardship: "They are already recognized on campus for preserving and providing access to other types of content; and they have strong links with the disciplinary communities.". In case of full text articles (published in scientific journals), books and conference proceedings, libraries nowadays rarely keep local copies of these publications. Instead, they offer access to remote network locations where library users can access and use scientific information resources. As a rule, libraries do not keep research data as part of their holdings. However, the times have changed, and universities would like libraries to become keepers of the local research output. For Angevaare (2009, p. 7) "Libraries have at least three crucial attributes which make them uniquely positioned to curate the output of academic research: 1. they have a mission that includes long-term preservation; 2. they have structural funding; 3. they have a network in the research community.". The question is whether the academic and research libraries are capable of data curation as this type of activity requires constant funding, trained library staff, adequate hardware and software, written data curation policies, logistics support from the academic institution to which library belongs etc. For some libraries, research data curation is an unwanted activity and they reluctantly accept this task. Joint (2007, p. 452) pointed out that „Scientists value their raw data more than the bibliographic expression of that data, and view the preservation of raw data as the prime curational challenge for the knowledge professions such as librarians and archivists.".

If libraries are not going to take care of research data, then we should think about new type of organizations which role would be providing support to e-science by integrating different research material in digital format including datasets of previous research. National Science Foundation (NSF, 2007.) calls for "(...) new type of organizations which will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise to: provide reliable digital data preservation, access, integration, and analysis capabilities over a decades1long timeline; continuously anticipate and adapt to changes in technologies and in user needs and expectations; engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and serve as component elements of an interoperable data preservation and access network.". In addition to university libraries, research data are kept by organizations which purpose is to support scientific research and to preserve the output of scientific endeavor. One such example is Digital curation centre

(http://www.dcc.ac.uk) in Great Britain. This organization is "(…) gateway to the technical solutions, curation tools and learning resources that can help data custodians like you to build capacity for digital curation." (Digital curation centre, 2010). These and other similar organizations (among which some are working for profit) will have a significant role in determining the future development of science in general. As McGovern and McKay (2008, p. 262) pointed out, the biggest problem will be finding organizations that "(…) can sustain their commitment to archive the digital content (…)".
These and other suggestions about who will take the responsibility for long term preservation of output of scientific research will help in discovering pros and cons of each possible solution offered. No final decision has been made yet about what organization will take the responsibility for data curation. At the moment, this responsibility is shared between scientists themselves, libraries and different organizations which curate data on commercial bases for profit. The principal question is who among them will demonstrate more dedication for preservation of research output in the long run?

Where research data are to be kept?

Digital curation needs reliable digital information systems that guarantee long term preservation, integrity and accessibility of research data. Digital repositories are most recent type of information resources that emerged in the academic community in the first half of the 1990s. Digital institutional repository (a digital information repository that is a part of university or other institution) is a digital archive of the intellectual product created by the faculty, research staff, and students of an institution and accessible to end users both within and outside of the institution, with few if any barriers to access (Johnson, 2002). An institutional digital repository can contain e-prints of scientific papers, research data, but also e-learning materials and other forms of institutional intellectual outputs, which are generally not published or preserved elsewhere (Hockx-Yu, 2006, pp. 234-235). Digital institutional repositories have a great importance in preservation of the scientific output: they focus organizational attention on managing digital content; they provide a potential entry point even a back door—for getting content into digital preservation programs; depositors and other stakeholders in institutional repositories may learn about digital preservation issues when they deposit digital content into institutional repositories; institutional repositories may offer an opportunity to address preservation planning priorities by providing guidelines and tools for depositors to prepare archive-ready digital content and they preserve retiring faculty's digital legacy (McGovern and McKay, 2008, p. 268). Digital repositories can store different file formats and types of content. Since data formats can become obsolete, repositories should only accept data in standardized formats. So far digital repositories have proved to be the best possible solution for the problem of data cu-

ration in this transitional period. Digital repositories are also the core of many projects related to data curation. Some examples of such projects are: Digital Curation Centre at http://www.dcc.ac.uk, DARIAH at http://www.dariah.eu/, University of Minnesota project at http://www.lib.umn.edu/datamanagement/archiving, National Geological and Geophysical Data Preservation Program at http://datapreservation.usgs.gov/index.shtml etc. There are also organisations like Datacite at http://datacite.org/ which help establish easier access to research data. Both projects and organisations help make understanding of data curation process better and help scholars and all other interested parties to get involved in this important activity.

Conclusion

"Maintenance of a complete and accurate scholarly record, including the portion in digital form, is essential for continued progress in research and learning." (Beagrie, Chruszcz, and Lavoie, 2008, p. 16). We can agree that this idea is one of the most important ideas that will support the development of science in the future by taking care of the output of its previous research so it can be integrated into new scientific research. As Isaac Newton wrote in 1676 "If I have seen a little further it is by standing on the shoulders of Giants." (The Phrase Finder, 2010). Today, a significant attention is given to the preservation of printed (as well as digital) scientific output such as journal articles, books, and conference papers published in proceeding, and less attention is given to the preservation of research data. Research data has important role in science as it can provide support to new scientific research. To preserve research data, a new type of activity called data curation has been developed. The purpose of this activity is to facilitate re-use of data of previously completed scientific research in a new scientific research. According to Van Horik (2008, p. 132) "Digital curation or data curation is needed to maintain digital materials, such as research data, over their entire life cycle and over time for current and future generations of users.". For this activity to be successful, science needs adequate infrastructure which will enable collecting, storing, organizing and re-use of research data. With development of that infrastructure, the science will be able to access more easily its fundaments and make new scientific discoveries possible.

References

Angevaare, Inge. Taking Care of Digital Collections and Data: 'Curation' and Organisational Choices for Research Libraries. Liber Quarterly. 19(2009), 1; 1-12.

Digital Curation Centre. http://www.dcc.ac.uk (3.6.2011.)

E-Science Data Curation. http://www.jisc.ac.uk/publications/generalpublications/2004/pub_escience.aspx. (1.5.2011.)

Harvey, Ross. Digital Curation : a how-to-do-it manual. New York, London : Neal-Schuman Publishers, 2010.

Hey, Tony; Trefethen, Anne. The Data Deluge: An e-Science Perspective. // Grid Computing—Making the Global Infrastructure a Reality. New York: John Wiley, 2003 http://eprints.ecs.soton.ac.uk/7648/1/The_Data_Deluge.pdf (5.4.2011.)

Hockx-Yu, Helen. Digital preservation in the context of institutional repositories. // Program: electronic library and information systems. 40(2006), 3; 234-245.

Ruusalepp, Raivo. Infrastructure Planning and Data Curation: A Comparative Study of International Approaches to Enabling the Sharing of Research Data. JISC, 2008. http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/Data-Sharing-Report.pdf (10.6.2011.)

Joint, Nicholas. Data preservation, the new science and the practitioner librarian. // Library Review. 56(2007), 6; 450-455.

Jones, Maggie. Digital Preservation Activities in the U.K – building the infrastructure. http://archive.ifla.org/IV/ifla69/papers/129e-Jones.pdf (9.2.2011.)

Johnson, Richard K. Institutional repositories: partnering with faculty to enhance scholarly communication. // D-Lib Magazine. 8(2002), 11, http://www.dlib.org/dlib/november02/ johnson/ (1.5.2011.)

Lyon, Liz. Dealing with data: roles, rights, responsibilities and relationships. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf (9.2.2011.)

Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. Arlington: National Science Foundation, 2005.

Lord, Philip.; Macdonald, Alison. Data curation for e-science in the UK: an audit to establish requirements for future curation and provision: e-Science Curation Report. http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf (9.2.2010.).

McGovern, Nancy Y.; McKay, Aprille C. Leveraging Short-term Opportunities to Address Long-term Obligations: A Perspective on Institutional Repositories and Digital Preservation Programs. // Library Trends. 57(2008), 2; 262-279.

Ogburn, Joyce L. The Imperative for Data Curation. // Libraries and the Academy. 10(2010), 2; 41–246.

Shearer, Kathleen; Argáez, Diego. Addressing the Research Data Gap: A Review of Novel Services for Libraries. Ottawa: CARL, 2010.

Shreeves, Sarah L.; Cragin, Melissa H. Introduction: Institutional Repositories: Current State and Future. // Library Trends. 57(2008),2 ; 89-97.

Sustainable Digital Data Preservation and Access Network Partners (DataNet), National Science Foundation, 2007 http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm (28.4.2011.)

Van Horik, Rene. Data curation. // A Driver's Guide to European Repositories. Amsterdam : Amsterdam University press, 2008. 131-152.

The Phrase Finder http://www.phrases.org.uk/meanings/268025.html (10.6.2011.)