




The 2011 Canadian Research Data Summit

Ottawa Convention Centre
September 14th - 15th, 2011

Mapping the Data Landscape

Table of Contents

1. Research, Reuse, and Innovation 3
2. The Stewardship of Research Data in Canada: A Gap Analysis 4
3. The Challenges of Data Sharing 6
4. International Models of Research Data Stewardship 7
5. The Power of Digital Data: Snapshots 10
6. Canadian Research Data Stewardship: A Vision for the Future 14
7. Appendices 15
   - *Glossary of Terms* 15
   - *Further Reading* 16

About the Summit

On September 14-15, 2011, The 2011 Canadian Research Data Summit brings together 100-150 senior researchers, high level policy makers, university administrators, and members of the private sector. Together, participants will work on formulating a shared strategy for addressing the challenges and opportunities for maximizing the benefits of our collective investment in research data in Canada. The Summit will act as a catalyst for the development of a made-in-Canada approach for maximizing the availability and use of research data.

About the Research Data Strategy Working Group

The Research Data Strategy Working Group is a collaborative effort launched in 2008 to address the challenges and issues surrounding the access and preservation of data arising from Canadian research. This multi-disciplinary group of universities, institutes, libraries, operators of research infrastructure, granting agencies, governments, and individual researchers are united through a shared recognition of the pressing need to deal with Canadian data stewardship issues. Together, they are focussing on the necessary actions, next steps and leadership roles that researchers and institutions can take to ensure Canada's research data are accessible and usable for current and future generations of researchers.

http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca

1. Research, Reuse, and Innovation

Digital data are revolutionizing the way research is being carried out, leading to a new data-centric way of thinking. Vast amounts of data are produced through publicly funded research in Canada, almost all of which are now in digital format. These data have virtually unlimited potential to be re-used in innovative ways- by researchers, industry, policy makers, and civil society- if they are properly managed in an infrastructure that provides long-term preservation and access. The impact of digital data is being felt far beyond the realm of research. A special report on big data published by the Economist states that "Data are becoming the new raw material of business: an economic input almost on a par with capital and labour." *(The Economist. Data, data, everywhere, Feb 25, 2010)*

"Like other essential factors of production such as hard assets and human capital, it is increasingly the case that much of modern economic activity, innovation and growth simply couldn't take place without data." *- (Mckinsey & Company. Big Data: The next frontier for innovation, competition, and productivity, March 2011)*

Data reuse leads to innovation.

Considerable advances in science and innovation have already been achieved through data sharing. For example, international domain archives such as Genbank have contributed to accelerated advances of discovery in genetics over the last decade. Similarly, a huge industry has flourished because of the widespread availability of geospatial data, including the development of products like GPS devices and Google maps that are used in people's everyday lives. The analysis of archived seismic data using new techniques has led to the discovery of new oil in previously "spent" wells. The ability to combine large health-related datasets with information from many other sources is enabling critical advances in public health. Imagine the impact if all research data were made available on a much broader scale!

"Managed well, data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account. Managed poorly, it can cause great harm."
(The Economist. Information: Making Use of the Data Deluge, June 2011)

To derive benefits, data must be managed.

As the volume of research data grows exponentially, so must the efforts to ensure that they are preserved, accessible, and understandable. For research data to be available for future use, long-term preservation must be a goal at the time data are created. Such data must be housed in an enduring institutional environment and accompanied by the necessary information (metadata) to ensure they are understandable. Furthermore, confidentiality of human subjects, intellectual property, and security of the data must be ensured where appropriate. All stakeholders in the research system, including governments, funding agencies, universities, disciplinary and specialist communities must be engaged in a coordinated and systematic way for this to be accomplished.

"Researchers and engineers need to have easy and quick access to the results of science. This is not a luxury for Europe, but a must, if we want to compete globally."
- (Neelie Kroes, Vice-President, European Commission. June 29, 2011)

The trend is truly global.

Many countries are making major investments in infrastructure to support the management, preservation and sharing of research data. Australia, for example, now treats research data as a strategic national resource and is carefully managing their research data accordingly. Research funders around the world are implementing policies to ensure that the data created through their funded projects are available for review and reuse. Recently, a group of major international funders of health research, including the National Institutes of Health in the United States and the Canadian Institutes of Health Research, adopted a set of principles on sharing research data with the intention of increasing the availability of data to accelerate advances in public health.

A coordinated approach is needed.

A coordinated and national approach to managing research data in Canada is required if we are to gain greater and longer term benefits, both socially and economically, from the extensive public investments that are made in research. A series of national consultations over the last decade has called for a national approach to managing Canada's research data assets. In addition, in 2004, Canada, along with 33 other nations, agreed to the OECD Principles and Guidelines for Access to Research Data from Public Funding. The federal government has launched an open government data initiative aimed at expanding the creative use of government-generated data into the non-governmental sphere and has recently committed to developing a policy that will strengthen and standardize the management and stewardship of all federal data; and some Canadian researchers are already involved in projects with leading-edge data management procedures. However, these initiatives represent only a small portion of the publicly-funded research data produced in Canada.

Canada must act now.

The way that we, as a nation, choose to manage our research data will directly impact our ability to participate in the evolving global data-intensive research environment. The absence of agreed upon data stewardship strategy could place Canada at a disadvantage compared to several of our international competitors. A coordinated and national approach to managing research data is needed to ensure that Canadians and others derive greater and more long-term benefit, both socially and economically, from the extensive public investments being made in research.

2. Stewardship of Research Data in Canada

A Gap Analysis

Data are increasingly being recognized as research assets with value beyond the original purpose for which they were intended. Numerous options exist to use and exploit these resources in ways to discover new knowledge and provide Canada with a competitive edge.

In order to capitalize on the opportunity to further exploit research data in Canada, we must actively conserve these digital resources. The purpose of this Gap Analysis is to identify the specific areas that need to be addressed to ensure that Canada effectively manages its research data. The analysis was originally produced in 2008 and updated again in 2011.

The Data Stewardship Lifecycle

The data stewardship lifecycle illustrates the various stages through which data travel as they are collected, used, preserved and reused, ensuring their value is maximized. A basic 4-stage model of the data lifecycle provides a useful framework for analyzing the state of data stewardship in Canada.

DATA PRODUCTION: This stage includes all activities involved in the planning, collecting, processing,analysis and maintenance of data in the original research project.

DATA DISSEMINATION: This stage involves the preparation of data for use by others and the establishment of procedures, documentation and methods for disseminating data.

LONG-TERM DATA MANAGEMENT: This stage involves activities supportive of the preservation and long-term access to research data.

DATA DISCOVERY AND REPURPOSING: This stage involves mechanisms and activities to enable the discovery and re-use of data.

In order for data to be used to their potential, they must be actively managed throughout **all four stages** of the lifecycle!

Good metadata is crucial for data stewardship, and different types of metadata are required as data travel through the lifecycle.

Ideal States

The analysis examined the difference between ideal states and current states across 10 indicators of data stewardship. The ideal states for each indicator are described here.

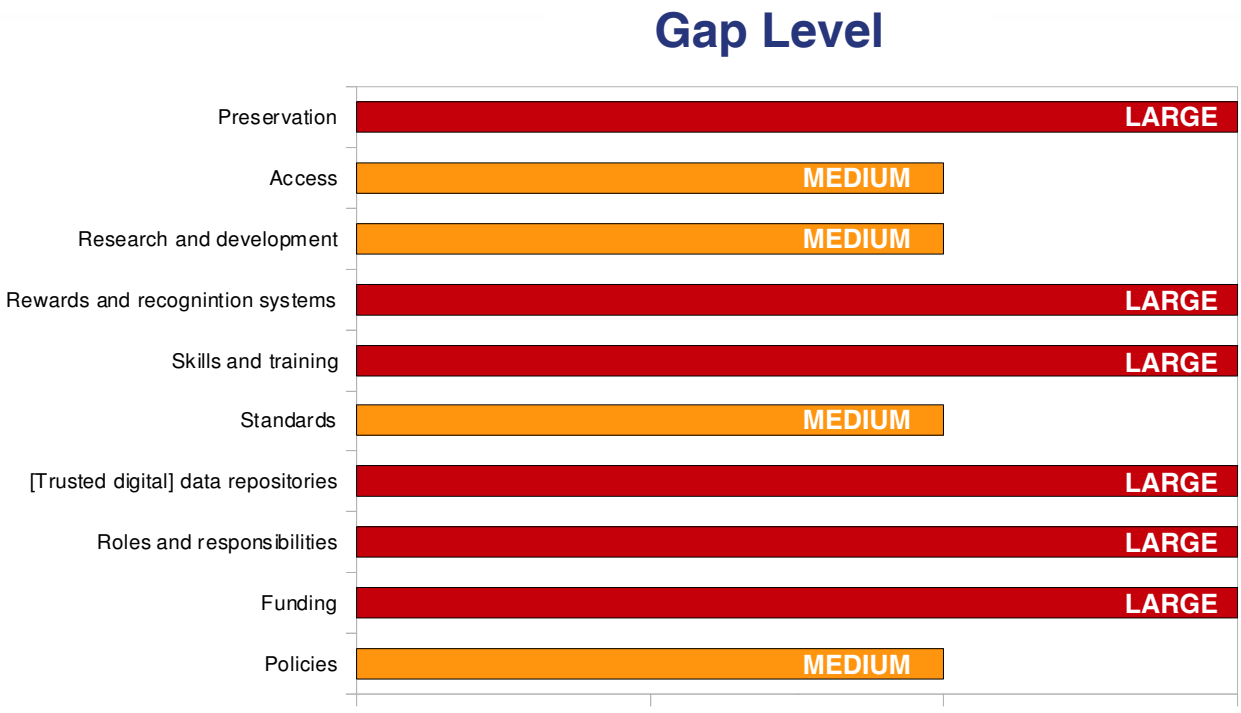
Indicator	Ideal State
Policies	Canadian organizations have coherent and cohesive policies that govern the management of data across disciplines.
Funding	Together, the range of funding mechanism cover costs throughout the data lifecycle, including the preservation of research data.
Roles and responsibilities	Each stakeholder in the data lifecycle has a distinct set of responsibilities, and works together to pursue higher-level stewardship goals; and *Canada has a national entity serving as a focal point to coordinate data management and stewardship activities across the country.*
Data repositories	Canada has a comprehensive network of trusted and interoperable data repositories.
Standards	There is widespread adherence to standards throughout the data lifecycle, including metadata standards.
Skills and training	Data stewardship activities are supported by specially trained data and information professionals, and researchers are well educated on the principles of data stewardship.
Rewards and recognition	Reward systems for researchers widely recognize contributions to research data management.
Research and development	Canada has a coordinated approach to R&D activities in support of data stewardship needs.
Access	There is widespread access to publicly funded research data, with appropriate mechanisms and licenses in place for articulating the terms of access where appropriate.
Preservation	Research data with enduring value are preserved using standards-based, active management practices throughout the data lifecycle.

Results

Although there are significant differences across disciplines, the analysis found considerable gaps between our current environment and the ideal state of data stewardship in Canada.

The existing piecemeal approach has resulted in serious gaps throughout the data stewardship lifecycle. In particular, Canada lacks the policies, infrastructure, and funding mechanisms to ensure that there is a continuum of care and respect for research data to ensure their value beyond the lifespan of the research project. In addition, in many disciplines, there exists a research culture that does not recognize the value of research data management. Large amounts of data are rendered inaccessible because of the absence of services and procedures to deliver data to other researchers. There are few funding mechanisms for maintaining data beyond the lifespan of a research project, nor are there clear lines of responsibilities for managing research data. Furthermore, the infrastructure to support data preservation and access, such as data repositories, is inadequate.

There are significant risks associated with doing nothing. Canadian researchers will not have the tools they need to participate in leading-edge research. The lack of a strategic, national vision for managing research data in Canada will have an impact on our innovation capacity.



Closing the Gaps

In order for Canada to remain competitive, we must develop a strategy for data stewardship. Given the scope of the challenge, it is clear that these issues cannot be resolved in isolation. Canada's governments, universities and research communities must work together to address the gaps and ensure that Canada remains competitive. The strategy should involve a multi-pronged approach that builds on the foundation already beginning to develop through other projects in Canada (e.g. International Polar Year Data Assembly Network, Research Data Centres, Neptune Canada, Open Data Government Initiative, Text Analysis Portal for Research, etc.).

The strategy should:

- Expand on existing policies that govern the management of research data
- Invest further in Canada's data repository infrastructure
- Create more training opportunities for data scientists and researchers, and
- Include an entity to ensure data stewardship activities across organizations and disciplines in Canada are coordinated.

The full report of the Gap Analysis is available on the Research Data Canada website.

3. The Challenges of Data Sharing

"But there are many challenges. How can we organise such a fiendishly complicated global effort, without hindering its flexibility and openness? How do we incentivise researchers, companies, and individuals to contribute their own data to the e-infrastructure – while still trusting that they can protect their privacy or ownership? How can we manage to preserve all this data, despite changing technologies and needs? How to convey the context and provenance of the data? How to pay for it all?"

(Riding the Wave. Final report of the High level Expert Group on Scientific Data. European Commission. October 2010)

Below is a list of some of the recognized challenges involved in achieving widespread data sharing and reuse in Canada:

- Managing increasingly large volumes of data
- Funding for infrastructure and services
- Ensuring confidentiality of research data about human subjects
- Protecting intellectual property
- Changing research culture(s) to recognize the value of data sharing
- Educating and recruiting skilled data scientists
- Training for researchers in managing data
- Ensuring interoperability across disciplines and internationally
- Technological complexity of data management systems
- Coordinating independent initiatives to develop a national data stewardship system

4. International Models of Research Data Stewardship

Many jurisdictions are making major investments in systems to support the management and sharing of research data. Globally, the value of reusing research data has been recognized as essential for advancements in science and innovation. The examples below highlight a variety of approaches to data stewardship being implemented internationally.

Research is a global endeavour and national data stewardship frameworks must ensure interoperability of systems across national boundaries.

Australia

In 2008, Australia launched a comprehensive national program for data sharing called the Australian National Data Service (ANDS) as part of its National Collaborative Research Infrastructure Strategy. The aim of ANDS is to create a national infrastructure that will enable Australian researchers to easily deposit, discover, access and re-use research data. The Government of Australia is investing heavily in this initiative. In addition to a $24 million budget for the years 2007 to 2011, ANDS was allocated an additional $48 million for 2009-2011 to create and develop an Australian Research Data Commons research infrastructure.

The long term (10-year) objectives for Australian National Data Service are to transform collections of Australian research data into a cohesive network of research repositories; assist Australian research data managers to become experts in creating, managing and sharing research data under well formed and maintained data management policies; increase the amount of research data that is routinely deposited into stable, accessible and sustainable data management and preservation environments; and more.

The ANDS approach is to work with Australia's research institutions to build capacity at the local level so that they have the infrastructure and expertise to support the capture, management, sharing, and dissemination of data. ANDS then connects those collections so that they can be found and used through the central service called the Australian Research Data Commons.

United States

In the US, there have been several activities related to data stewardship over the last several years. Open Data is a project that was launched in 2010 which aims to increase public access to high value, machine readable data sets generated by the Executive Branch of the Federal Government. In addition, the federal government recently convened a Big Data Senior Steering Group to identify current big data research and development activities across the Federal government. This initiative offers opportunities for coordination, and identification of concerns and requirements for a national-level effort. It also aims to develop strategies to leverage current investments and resources in data management.

In the research sector, both the National Science Foundation (NSF) and the National Institutes of Health (NIH) have implemented policies requiring researchers to make their data available at the end of a research project. The NSF also requires researchers to include "data management plans" in their funding proposals, as does the NIH for projects with budgets exceeding $500,000. According to the NSF, "This is the first step in what will be a more comprehensive approach to data policy."

In 2005, the NSF instituted an Office for Cyberinfrastructure (OCI), to serve as a catalyst in developing a system of science and engineering data collections that is open, extensible, and evolvable. This Office also supports the development of a new generation of tools and services for data discovery, integration, visualization, analysis and preservation. To realize this vision, NSF has provided $100 million in funding for a program called Sustainable Digital Data Preservation and Access Network Partners (DataNet). The program is working with some of the large scale data repositories to develop "new methods, management structures and technologies to manage the diversity, size, and complexity of current and future data sets and data streams by creating a set of exemplar national and global data research infrastructure organizations."

European Commission

In October 2010, a High Level Expert Group on Scientific Data submitted a report to the European Commission containing recommendations about how Europe can gain from the rising tide of scientific data. Their vision is to develop a global framework for research data in which all countries can participate, through developing complementary and interoperable data management systems. The report goes on to assert that "The European Union has an important, coordinating role in achieving this vision – through its Digital Agenda, its Framework Programme and the policies embodied in its European Research Area initiatives. Equally, there is the opportunity for the EU institutions to lead in creating a common, world-wide vision."

GRDI2020: Towards a 10-Year Vision for Global Research Data Infrastructures, is a initiative that seeks to define a roadmap for evolving global scientific data infrastructures. The preliminary report envisions the establishment of Interoperable Science Ecosystems, composed of Digital Data Libraries, Digital Data Archives, and Digital Research Libraries and Global Scientific Data Infrastructures which will act as the enablers of these interoperable science ecosystems. The final report will be made available in 2012 and will contain recommendations for member states in terms of implementation.

The EC has recognized that for the EU to excel in research, researchers need to have easy and quick access to the results of science. To facilitate this, Neelie Kroes, the Digital Agenda Commissioner for the European Commission, recently announced that the EC will be expanding their open access policies to require that the data resulting from EU-funded research to be made widely available.

United Kingdom

The UK already has a very robust infrastructure of national repositories of research data in a variety of disciplines. The repositories are managed by several of the Research Councils UK (RCUK) and both collect the data created through research and provide access to this data for others. The RCUK has also published a set of common principles on which their research data policies will be based. These Principles state, "Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property."

The UK Data Forum has published the UK Strategy for Data Resources for Social and Economic Research 2009-2012, which sets out priorities for the development of research data resources both within the social sciences and at the boundaries between the social sciences and other areas of scientific enquiry. It is an action plan that presents a coordinated planned approach to funding, development and maintenance of data required for research in the social sciences.

In addition, another funding agency in the UK, the Joint Information Systems Committee (JISC), is funding a number of data related activities at UK universities through its Managing Research Data Programme. The program aims to build capacity for managing research data through development of data management infrastructure, training and skills, and tools for managing research data. JISC also funds the Digital Curation Centre (DCC), a national centre of expertise for curating digital research data. In addition to providing expert advice and training to researchers in the area of data management, the DCC is a gateway to the technical solutions, curation tools and learning resources that help data custodians build capacity for digital curation.

Germany

In 2010 the Alliance of German Science Organizations adopted "Principles for the Handling of Research Data". The principles state that research data should be freely accessible, they should be easy to get and they should be professionally curated on a long-term basis. In addition, in 2010, the Alliance added additional requirements for grant proposals, stating that if research projects are producing scientific data with potential re-use value, they must describe in their proposal how these data will be securely stored and made available for other scientists."

Germany is the home to a number of the World Data Centres, which collect solar, geophysical and related environmental data, in addition to a number of other discipline-based data repositories. In order to expand their capacity to better manage the data produced through research, the German Research Foundation (DFG), has recently funded 27 projects to develop the research data management infrastructure in Germany. The projects deal with mid- to long-term archiving, discipline-specific metadata and systems for linking publications with research data, as well as with questions of quality control.

The German National Library of Science and Technology also houses DataCite, a global organization which includes the Canada Institute for Scientific and Technical Information as the Canadian member. DataCite aims to develop an infrastructure that supports simple and effective methods of data citation, discovery, and access; initially by working with data centres to assign persistent identifiers to data sets.

Summary table of national action on data stewardship

	Australia	Canada	European Commission	Germany	United Kingdom	United States
Policies	x	x	xx	xx	xxx	x
Infrastructure	xxx	x	x	xx	xxx	xx
Skills and training	xx		x	x	xxx	x
Funding	xxx	x	xx	xx	xxx	xxx
Coordination	xxx		x	xx	xx	x

5. The Power of Digital Data: *Snapshots*

Considerable advances have already been achieved through data stewardship and sharing. These short snapshots illustrate the vast potential of digital research data for reuse, exploitation and innovation.

Archived data leads to Canadian discovery of neutron star

University of Alberta physics professor Craig Heinke has solved a mystery that lies 11,000 light years beyond Earth and has puzzled astronomers for years. When a supernova, an exploding star 20 times heavier than our sun, blasted apart, it left behind a small core, a 20 kilometre-wide remnant, which Heinke and his colleague, Wynn Ho, identified as a neutron star. It's the youngest neutron star ever identified, and its atmosphere, a thin layer of carbon, is one of a kind. Heinke's breakthrough reveals part of a neutron star's life cycle that had never been seen before.

Heinke's discovery was based on archived data from NASA's Chandra X-ray Observatory satellite According to Dr. Heinke, "It's been many years since I have written a paper that did *not* use archived X-ray or Hubble data, usually along with new data (but not always). Astronomers are increasingly using archived data to make new discoveries. There are so many wavelength bands in which objects can be studied by different telescopes, and the same data can be used for various purposes by different people." According to the Space Telescope Science Institute, there are now more refereed papers each year based on their archived data, than newly collected data.

New cancer database helps physicians share medical treatments

In 2010, Georgetown Lombardi Comprehensive Cancer Center launched the Georgetown Database of Cancer or G-DOC. Under development for two years, G-DOC is a repository for biological information that normally would only be available in scattered information libraries and tissue banks, if at all. The overall goal of G-DOC is to accelerate the ability of physicians to tailor or personalize medical treatment for patients. For physicians and researchers, G-DOC means they need only access a single website to open a unique and comprehensive information resource on cancer interventions.

Before G-DOC, there was no way of connecting treatment outcomes across a wide variety of indicators. The database facilitates the analysis of genes, proteins and other molecules from tumour tissue, as well as detailed outcome data collected, to help predict, for example, who is likely to be cured with surgery alone, and who may need more aggressive treatment. G-DOC enables physicians to choose the best form of treatment for their individual patients, improving outcomes and saving lives.

Open government data drives innovation

Like many other countries, Canadian governments (municipal, provincial and federal) have started making some of the data they produce available over the internet. The data is released in machine-readable form and allows for mash-ups and driving the creation of new tools and products. Innovators across the world are responding. Companies including Google, Microsoft, and smaller hi-tech companies as well as individuals have begun to build amazing apps that make use of this data. The release of transportation data, for example, has led to the development of numerous public applications that allow users to track public transport or traffic jams in real time. The City of Toronto recently launched Wellbeing Toronto, giving users the ability to map 140 officially-defined neighbourhoods by dozens of different data points, from arson to breast cancer screening to sports facilities. As the number of data sets available through governments grows, the potential for innovation becomes virtually limitless. Currently, through the open data portal, the Government of Canada has made available over 800 general data sets and over 200,000 geospatial data sets.

Harnessing the power of citizen science

'Citizen science' refers to scientific projects in which individuals or networks of citizen volunteers perform or manage research-related tasks such as observation, measurement, or computation. The use of citizen-science networks allows scientists to accomplish research objectives more feasibly than would otherwise be possible. In addition, these projects promote public engagement with the research, as well as with science in general. One example of citizen science in action is the Old Weather Project. The project is managed by Oxford University in the UK, but involves collaboration between a diverse collection of people and institutions across the world. The Old Weather Project involves analyzing historical weather data in order to test the validity of models of the Earth's climate. If scientists can correctly account for what the weather was doing in the past, then they can have more confidence in their predictions of the future.

The project is asking citizens to transcribe hundreds of thousands of digitized images of weather data from naval log books into machine readable text. The data will be used to fill in gaps in large-scale climate models. The logs come from British naval ships that sailed from 1905 to 1929, a time when sailors wrote down temperature, wind and other climate data every four hours. The digitized logs cannot be processed using computers, as they are written in often difficult to decipher hand writing. The project is analyzing logs from over 500 ships and has enlisted tens of thousands of volunteers.

History unveiled through mapping spatial information

According to the New York Times, humanists now have a new tool that can help them better visualize historical events. A new generation of digital maps has given rise to an academic field known as spatial humanities. Historians, literary theorists, archaeologists and others are using Geographic Information Systems to re-examine real and fictional places. "Mapping spatial information reveals part of human history that otherwise we couldn't possibly know," said Anne Kelly Knowles, a geographer at Middlebury College in Vermont. "It enables you to see patterns and information that are literally invisible." It adds layers of information to a map that can be added or taken off at will in various combinations; the same location can also be viewed back and forth over time at the click of a mouse.

New methods of computer-assisted geographic analysis can also offer new interpretations of familiar topics. Geoff Cunfer, a historian at the University of Saskatchewan, revisited causes of the 1930s Dust Bowl by analyzing data from all 208 counties in Texas, New Mexico, Colorado, Oklahoma and Kansas that were affected. Using reports of annual precipitation, unplowed grassland, wind direction, droughts, agricultural censuses, historical studies and previous reports on dust storms — "a messy shoebox full of newspaper clippings" — Dr. Cunfer created data sets that could be plotted on maps. He found that the traditional explanation of farmers' extensively plowing the land without care for environmental limits was only true in some places. Barely plowed Southern counties also suffered from the plague of dust.

Recreating data can be very expensive

Many organizations find out the hard way that the costs of maintaining data properly are insignificant compared to the expenses incurred in trying to replace or recreate lost data. For example, data from the Canada Land inventory, a comprehensive multi-disciplinary land inventory of rural Canada, was very nearly lost due to neglect.

During the 1960s and 1970s, Canada expended much time and effort attempting to define the location of its natural resources and the geographic extent of human activity on rural and urban lands. Two projects of particular interest were the Canada Land Inventory (CLI) that mapped land use, and the Canada Land Use Monitoring Program (CLUMP) that looked at land use change. Together these represented one of the earliest and largest investments in constructing GIS databases ever undertaken. In 1995 use of the Canada Land Data System was discontinued. This decision placed the data, which had been collected through the use of tens of millions of Canadian tax payers' dollars, at risk of being lost to the great electronic void.

It took four federal government departments over 5 years and many thousands of dollars to restore the data. It would have been nearly impossible to recreate this valuable data that has since sparked a number of new initiatives.

Hundreds of researchers re-use childhood development survey data

In 2007, the Journal of Applied Developmental Psychology published a special issue to highlight how data from the Study of Early Child Care and Youth Development had been used by researchers to address a range of research questions not envisioned in the original study plan. The data sets contain longitudinal data about family, day care and school environments of over 1000 children, tracked from birth through the age of fifteen. Most of the publications by original study investigators focused on child-care issues. However, the editors found that the original data have been used by hundreds of other investigators studying a wide range of other research problems, such as child development, women's employment patterns, quality of academic instruction, family social and psychological processes, and more.

Sharing data leads to early diagnosis of Alzheimer's disease

In 2003, a group of scientists from the National Institutes of Health, the Food and Drug Administration, the drug and medical-imaging industries, universities, and nonprofit groups joined in a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain. Prior to this project, scientists were looking for biomarkers, but they were not getting very far because many different scientists in different universities were doing their own research with their own patients and using their own methods. The way they were documenting biomarkers was diverse and there was little exchange of data across different research projects. In order to address these barriers, they came to an agreement to share all their data, making every single finding immediately available to anyone with a computer anywhere in the world. Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

Data sharing improves transparency in the climate sciences

Climate scientists are under pressure to make their data, and their methods, more openly available, both to fellow scientists and the public. Political pressure from sceptics is increasing public demands for access to code and data. Advocates for greater transparency argue that data sharing will make the science replicable and will improve its overall quality. Confidence in research findings depends upon the underpinning data and methods being open, reusable, and verifiable.

By traditional measures climate science is already open. Researchers publish their results in peer-reviewed journals; they share data with one another; they present at conferences and collaborate on projects. However, with new technology a whole new level of openness is possible. Data and code can be stored online in repositories and shared at the click of a mouse. This is leading to a growing number of research agencies releasing their data to the public and there are now terabytes of data available through international climate data centres for anyone who wants to download it.

Charles Darwin's personal library online

Charles Darwin's personal scientific library, the majority of which is held at Cambridge University Library, has been digitised in a collaborative effort involving Cambridge and other partners. The pages of Darwin's Library, smothered as they are in his tantalizing scrawl that give readers a direct insight into of the great Darwinian intellect. In total, Darwin's library amounted to 1,480 books, of which 730 contain abundant research notes in their margins. These annotated books are now being digitized and made available over the internet via the Biodiversity Heritage Library.

For scholars, having direct online access to these types of collections is a tremendous boon. It allows them to access collections previously not readily available; and it also enables them to make use of digital tools such as text analysis or digital annotation to assist with their research. According to the university librarian Cambridge University, The Darwin collections are among the most important and popular held in library. However, until his papers were digitized and put online, interested researchers would have to make the trip in person to the Cambridge Library.

Open ocean data from the NEPTUNE Project

NEPTUNE Canada at University of Victoria is collecting huge volumes of real-time data using cables that are laid on the ocean floor. The project represents a new approach to ocean science. NEPTUNE has an open data policy, and makes the data they collect available real time to everyone via their website. This approach is in contrast with a prevailing attitude by researchers that don't want to share data because they are worried about being "scooped". NEPTUNE data are used by researchers in many disciplines: geoscientists, ocean chemists, biologists, etc. The aim is to ensure that this data remains available and accessible over a long period of time so that researchers can measure effects of climate changes, warming, etc. The lifespan of the project is 25 years, however, there remains a lingering question about what will happen to the data once the project is over.

Analyzing such a large volume of data is a big challenge and NEPTUNE is looking at approaches that involve citizen scientists. They have recently received funding from Canada's Advanced Research and Innovation Network (CANARIE Inc.) to embark on a new project, "Data from the Deep, Judgment from the Crowd". This project will make use of crowdsourcing to engage large numbers of volunteers to help analyze some of the data collected through the project that are not easily decipherable by computer methods.

Sharing of genomics data accelerates progress

GenBank is an open access database of nucleic acid sequences that was launched in 1982 by the National Institutes of Health (NIH). When scientists first began sequencing proteins and DNA, it was an expensive and time consuming process, leading researchers to limit their sequencing to those genes and proteins for which they had a particular interest. A small number of groups began collecting sequencing data and would sometimes do comparisons that led to serendipitous discoveries. Eventually, a consensus emerged about the need for an international computer database which would allow researchers to systematically share data with each other, which led to the creation of GenBank.

Because of GenBank, researchers around the world were able to collaborate to map the human genome. The Human Genome Project opened its databases to the public in 1990. By 2003, it had succeeded in sequencing all the base pairs in the human genome and it did this under budget and more than two years ahead of schedule. Since 2003, GenBank has grown exponentially and now contains data on about 110 million sequences and 200 billion base pairs.

6. Canadian Research Data Stewardship

A Vision for the Future

We envision a future in which the value of our nation's investment in research is optimized and, as a result, new social-economic benefits are created for Canadians. As a consequence, there is broad societal recognition of the value of research data. Researchers from all disciplines have widespread access to research data enabling them to conduct leading-edge research in Canada and to participate actively in international data-intensive research endeavours. Industry, practitioners, and the public are able to exploit research data, where appropriate, for commercial, policy and creative purposes.

Research data in Canada are systematically managed, preserved and utilized to create knowledge, to stimulate innovation and to advance Canada's leadership in the global digital economy. Canada has a coordinated approach to managing research data that includes the following aspects:

Policies: Canadian organizations have established coherent and cohesive policies based on agreed-upon national data management principles that apply across disciplines and the lifecycle stages of research. Data management plans are an integral part of funded research and operationalize the goal of making data openly accessible. Furthermore, the contributions that researchers make in successfully producing, sharing and preserving data are acknowledged by institutional reward systems.

Roles and responsibilities: All stakeholders in the research process understand their roles in and perform their responsibilities across the data lifecycle. All stakeholders, through their distinct set of responsibilities, also act in partnership with other stakeholders to pursue higher-level stewardship goals collectively important to the entire research community.

Skills and training: Collectively, researchers, trained data scientists and information professionals provide the necessary skills to select, manage, preserve and provide access to research data. All other stakeholders in the research process are similarly well educated on their own roles and responsibilities with respect to data management and stewardship.

Infrastructure: Canada supports a national, collaborative, interoperable network of institutional services, including data repositories, data centres, data warehouses and data libraries, together with a high bandwidth research network that enables access. Collectively, these services collect, preserve and disseminate valued research data and are managed by a variety of organizations across sectors, such as governments, universities and others. Interoperability is achieved through metadata and adherence to standards and best practices.

Sustainability: Research data are entrusted to an enduring institutional environment with long-term commitments to preserving and providing access to such data. These institutions employ international and national data standards. Collectively, the range of funding mechanisms covers operational costs throughout the data lifecycle.

National coordination mechanism: Canada has a national entity serving as a focal point to support data management and stewardship activities across the country. Bringing together regional and disciplinary networks, this entity: (1) ensures a Canadian presence in international research data initiatives; (2) houses a training and resource centre dedicated to advancing research data skills, standards and practices; and (3) helps provide relevant policy advice along with other organizations.

7. Appendices

Glossary of Terms

Data curation: The activity of managing the use of data from the point of creation to ensure they are available for discovery and re-use in the future. This management process includes storage and security of data, quality management, recording important information about the data, including its source, analysis methods and changes to the data, and preserving the data so that they can be accessed and re-used in the future.

Data interoperability: the structuring of data in such a way that diverse data sets can be integrated.

Data management: A process involving a broad range of activities for handling data, from administrative to technical aspects.

Data management plan: A formal document that outlines how a researcher or research project will handle their data both during their research, and after the project is completed.

Data policy: A set of high-level principles that establish a guiding framework for data management. A data policy can be used to address strategic aspects such as data access, relevant legal matters, data stewardship issues and custodial duties, data acquisition, and other issues.

Data science: A set of high-level principles that establish a guiding framework for data management. A data policy can be used to address strategic aspects such as data access, relevant legal matters, data stewardship issues and custodial duties, data acquisition, and other issues.

Data set: Any organized collection of data.

Data stewardship: An organizational plan of the roles and responsibilities of those overseeing the management of data across all stages of the data lifecycle, including its preservation. A large research project may involve several data stewards as the data moves from stage to stage across the lifecycle, while data stewardship in a small project may fall primarily upon the principal investigator and the organization taking responsibility for the preservation of the data.

Digital Object Identifier (DOI): A digital identifier for any object of intellectual property, in this case a data set. A DOI provides a means of persistently identifying a piece of intellectual property on a digital network and associating it with related current data in a structured extensible way.

Metadata: “Data about data”. Metadata is structured data which describes the characteristics of a resource, such as research data.

Research data: The factual records used as primary sources for research, and that are commonly accepted in the research community as necessary to validate research findings. Virtually all types of digital information have the potential to be research data, if they are being used as a primary resource for research.

Further Reading

Big Data: The next frontier for innovation, competition, and productivity, McKinsey & Company. March 2011

Cyberinfrastructure Vision for 21st Century Discovery, National Science Foundation. March 2007

Data, data, everywhere, The Economist. Feb 25, 2010

Data Sharing, Special. Nature. September 9, 2009.

Dealing with Data: Special Online Collection, Science Magazine. February 2011

Final Report: National Consultation on Access to Scientific Research Data (NCASRD), National Research Council, Ottawa. 2005

Harnessing the Power of Digital Data for Science and Society, Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. January 2009

National Science Foundation, Advisory Committee for Cyberinfrastructure, Task Force on Data and Visualization, Final Report, March 2011.

Research Data: Unseen Opportunities: An Awareness Toolkit commissioned by the Canadian Association of Research Libraries (CARL), CARL. 2010

Riding the Wave: How Europe can gain from the rising tide of scientific data, Final report of the High Level Expert Group on Scientific Data A submission to the European Commission. October 2010

Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information, Blue Ribbon Task Force on Sustainable Digital Preservation and Access. February 2010

Research Data Strategy Working Group

http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca

The 2011 Canadian Research Data Summit: Mapping the Data Landscape

Links and references from the digital version available at:
http://www.verney.ca/nrc2011/background.php

Page 2: Title Page

The 2011 Canadian Research Data Summit: Mapping the Data Landscape: http://www.verney.ca/nrc2011/
The Research Data Strategy Working Group: http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/

Page 3: Research, Reuse, and Innovation

The Economist, Data, data, everywhere, Feb 25, 2010: http://www.economist.com/node/15557443
McKinsey & Company, Big Data: The next frontier for innovation, competition, and productivity, March 2011: http://www.mckinsey.com/mgi/publications/big_data/index.asp
Statistics Canada, Surveying and Mapping Services. 2009: http://www.statcan.gc.ca/pub/63-254-x/2011001/part-partie1-eng.htm
The Economist, Information: Making Use of the Data Deluge, June 2011: http://ideas.economist.com/event/information/programme
Neelie Kroes, Vice-President, European Commission, June 29, 2011: www.youtube.com/watch?v=GIU14-3hYto

Page 4: The Stewardship of Research Data in Canada: A Gap Analysis

Stewardship of Research Data in Canada: Gap Analysis. Research Data Strategy Working Group, 2008

Page 6: The Challenges of Data Sharing

Riding the Wave. Final report of the High level Expert Group on Scientific Data. European Commission. October 2010: http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707

Page 7: International Models of Research Data Stewardship

Australia

Australian National Data Service (ANDS): http://www.ands.org.au/

United States

Data.Gov: http://www.data.gov/
Big Data Senior Steering Group: http://www.nitrd.gov/subcommittee/bigdata.aspx
Sustainable Digital Data Preservation and Access Network Partners: http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

European Commission

Riding the Wave. Final report of the High level Expert Group on Scientific Data. European Commission. October 2010: http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707
GRDI2020: Towards a 10-Year Vision for Global Research Data Infrastructures: http://www.grdi2020.eu/
GRDI2020 Preliminary Roadmap Report. Global Scientific Data Infrastructures: The Big Data Challenges: http://www.grdi2020.eu/Pages/SelectedDocument.aspx
Neelie Kroes, Vice-President, European Commission, June 29, 2011: www.youtube.com/watch?v=GIU14-3hYto

United Kingdom
RCUK Common Principles on Data Policy: http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
UK Strategy for Data Resources for Social and Economic Research 2009-2012: http://www.esrc.ac.uk/funding-and-guidance/tools-and-resources/research-resources/data-services/NDS/
Managing Research Data Programme: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx
Digital Curation Centre (DCC): http://www.dcc.ac.uk/

Germany
Principles for the Handling of Research Data: http://www.allianzinitiative.de/en/core_activities/research_data/principles/
World Data Centres: http://www.ngdc.noaa.gov/wdc/list.shtml
Development of Information Infrastructures for Research Data: DFG Approves 27 Projects: http://www.dfg.de/en/research_funding/announcements_proposals/info_wissenschaft_11_18/index.html
DataCite: http://datacite.org/

Page 10: The Power of Digital Data Snapshots

Archived data leads to Canadian discovery of neutron star
Brian Murphy. U of A physicist IDs young neutron star. Brian Murphy. University of Alberta International. Nov 4, 2009: http://www.uofaweb.ualberta.ca/uai_prospective/news.cfm?story=95435
Craig Heinke, Department of Physics, University of Alberta: http://www.ualberta.ca/~heinke/
Chandra X-ray Observatory: http://www.nasa.gov/mission_pages/chandra/main/index.html

New cancer database helps physicians share medical treatments
Georgetown Database of Cancer (G-DOC): https://gdoc.georgetown.edu/gdoc/

Open government data drives innovation
Wellbeing Toronto: http://www.toronto.ca/wellbeing/
Government of Canada, Open Data Pilot Project: http://www.data.gc.ca/

Harnessing the power of citizen science
Wikipedia. Citizen science: http://en.wikipedia.org/wiki/Citizen_science
Old Weather Project: http://www.oldweather.org/

History unveiled through mapping spatial information
Patricia Cohen. Digital Maps Are Giving Scholars the Historical Lay of the Land, New York Times. July 27, 2011: http://www.nytimes.com/2011/07/27/arts/geographic-information-systems-help-scholars-see-history.html?_r=1
Frontiers of Spatial Humanities (video): http://www.scholarslab.org/announcements/frontiers-in-spatial-humanities-video/
Geoff Cunfer, College of Arts and Science, University of Saskatchewan: http://artsandscience.usask.ca/profile/GCunfer

Recreating data can be very expensive
Author unknown. Back from the Brink: the story of the ressurection of the Canada Lans Inventory data.
Canada Land inventory: http://res.agr.ca/cansis/nsdb/cli/intro.html

Hundreds of researchers re-use childhood development survey data
Journal of Applied Developmental Psychology: http://www.sciencedirect.com/science/journal/01933973

Sharing data leads to early diagnosis of Alzheimer's disease
Gina Kolata. Sharing of Data Leads to Progress on Alzheimer's. New York Times. August 12, 2010: http://www.nytimes.com/2010/08/13/health/research/13alzheimer.html?_r=3

Data sharing improves transparency in the climate sciences
Kurt Kleiner. Data on demand. Nature Climate Change 29 March 2011: http://www.nature.com/nclimate/journal/v1/n1/full/nclimate1057.html#auth-1

Charles Darwin's personal library online
Charles Darwin's Library: http://www.biodiversitylibrary.org/collection/darwinlibrary

Open ocean data from the NEPTUNE Project
NEPTUNE Canada, University of Victoria: http://www.neptunecanada.ca/
Data from the Deep, Judgment from the Crowd: http://digitalfishers.net/?p=8#more-8

Sharing of genomics data accelerates progress
GenBank: http://www.ncbi.nlm.nih.gov/genbank/

Page 15: Glossary of Terms

Scholarly communications and curation: http://ands.org.au/guides/curation.continuum.html
What is research data? Australian National Data Service. http://ands.org.au/guides/what-is-research-data.pdf

le Sommet 2011 sur les données de recherche canadiennes:
Portrait de la situation des données

Les liens et adresse de la version numérique disponible à :
http://www.verney.ca/nrc2011/background.php

Page 2: Table des matières

le Sommet 2011 sur les données de recherche canadiennes: http://www.verney.ca/nrc2011/
Le Groupe de travail sur la stratégie des données de recherche: http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/

Page 3: Recherche, réutilisation et innovation

The Economist, Data, data, everywhere, Feb 25, 2010: http://www.economist.com/node/15557443
McKinsey & Company, Big Data: The next frontier for innovation, competition, and productivity, March 2011: http://www.mckinsey.com/mgi/publications/big_data/index.asp
Statistics Canada, Surveying and Mapping Services. 2009: http://www.statcan.gc.ca/pub/63-254-x/2011001/part-partie1-eng.htm
The Economist, Information: Making Use of the Data Deluge, June 2011: http://ideas.economist.com/event/information/programme
Neelie Kroes, Vice-President, European Commission, June 29, 2011: www.youtube.com/watch?v=GIU14-3hYto

Page 4: Gérance des données de recherche au Canada

Stewardship of Research Data in Canada: Gap Analysis. Research Data Strategy Working Group, 2008

Page 6: Défis liés à la mise en commun des données

Riding the Wave. Final report of the High level Expert Group on Scientific Data. European Commission. October 2010: http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707

Page 7: Modèles internationaux de gérance des données de recherche

Australie
Australian National Data Service (ANDS): http://www.ands.org.au/

États-Unis
Data.Gov: http://www.data.gov/
Big Data Senior Steering Group: http://www.nitrd.gov/subcommittee/bigdata.aspx
Sustainable Digital Data Preservation and Access Network Partners: http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

Commission européenne
Riding the Wave. Final report of the High level Expert Group on Scientific Data. European Commission. October 2010: http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707
GRDI2020: Towards a 10-Year Vision for Global Research Data Infrastructures: http://www.grdi2020.eu/
GRDI2020 Preliminary Roadmap Report. Global Scientific Data Infrastructures: The Big Data Challenges: http://www.grdi2020.eu/Pages/SelectedDocument.aspx
Neelie Kroes, Vice-President, European Commission, June 29, 2011: www.youtube.com/watch?v=GIU14-3hYto

Royaume-Uni
RCUK Common Principles on Data Policy: http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
UK Strategy for Data Resources for Social and Economic Research 2009-2012: http://www.esrc.ac.uk/funding-and-guidance/tools-and-resources/research-resources/data-services/NDS/
Managing Research Data Programme: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx
Digital Curation Centre (DCC): http://www.dcc.ac.uk/

Allemagne
Principles for the Handling of Research Data: http://www.allianzinitiative.de/en/core_activities/research_data/principles/
World Data Centres: http://www.ngdc.noaa.gov/wdc/list.shtml
Development of Information Infrastructures for Research Data: DFG Approves 27 Projects: http://www.dfg.de/en/research_funding/announcements_proposals/info_wissenschaft_11_18/index.html
DataCite: http://datacite.org/

Page 10: Pouvoir des données numériques: aperçus

Des données archivées mènent à la découverte canadienne de l'étoile à neutrons
Brian Murphy. U of A physicist IDs young neutron star. Brian Murphy. University of Alberta International. Nov 4, 2009: http://www.uofaweb.ualberta.ca/uai_prospective/news.cfm?story=95435
Craig Heinke, Department of Physics, University of Alberta: http://www.ualberta.ca/~heinke/
Chandra X-ray Observatory: http://www.nasa.gov/mission_pages/chandra/main/index.html

La nouvelle base de données sur le cancer aide les médecins à diffuser des traitements médicaux
Georgetown Database of Cancer (G-DOC): https://gdoc.georgetown.edu/gdoc/

Les données gouvernementales ouvertes stimulent l'innovation
Wellbeing Toronto: http://www.toronto.ca/wellbeing/
Government of Canada, Open Data Pilot Project: http://www.data.gc.ca/

Exploiter pleinement les possibilités qu'offre la science citoyenne
Wikipedia. Citizen science: http://en.wikipedia.org/wiki/Citizen_science
Old Weather Project: http://www.oldweather.org/

L'histoire reconstituée grâce au mappage de l'information spatiale
Patricia Cohen. Digital Maps Are Giving Scholars the Historical Lay of the Land, New York Times. July 27, 2011: http://www.nytimes.com/2011/07/27/arts/geographic-information-systems-help-scholars-see-history.html?_r=1
Frontiers of Spatial Humanities (video): http://www.scholarslab.org/announcements/frontiers-in-spatial-humanities-video/
Geoff Cunfer, College of Arts and Science, University of Saskatchewan: http://artsandscience.usask.ca/profile/GCunfer

La reconstitution des données peut être très coûteuse
Author unknown. Back from the Brink: the story of the ressurection of the Canada Lans Inventory data. Canada Land inventory: http://res.agr.ca/cansis/nsdb/cli/intro.html

Des centaines de chercheurs réutilisent les données des études sur le développement des enfants
Journal of Applied Developmental Psychology: http://www.sciencedirect.com/science/journal/01933973

La mise en commun des données mène au diagnostic précoce de la maladie d'Alzheimer
Gina Kolata. Sharing of Data Leads to Progress on Alzheimer's. New York Times. August 12, 2010: http://www.nytimes.com/2010/08/13/health/research/13alzheimer.html?_r=3

La mise en commun des données améliore la transparence dans les sciences climatiques
Kurt Kleiner. Data on demand. Nature Climate Change 29 March 2011: http://www.nature.com/nclimate/journal/v1/n1/full/nclimate1057.html#auth-1

La bibliothèque personnelle de Charles Darwin en ligne
Charles Darwin's Library: http://www.biodiversitylibrary.org/collection/darwinlibrary

Données océaniques ouvertes provenant du projet NEPTUNE
NEPTUNE Canada, University of Victoria: http://www.neptunecanada.ca/
Data from the Deep, Judgment from the Crowd: http://digitalfishers.net/?p=8#more-8

La mise en commun des données génomiques accélère les avancées
GenBank: http://www.ncbi.nlm.nih.gov/genbank/

Page 15: Glossaire

Scholarly communications and curation: http://ands.org.au/guides/curation.continuum.html
What is research data? Australian National Data Service. http://ands.org.au/guides/what-is-research-data.pdf