Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada

Consultation Document
(16 October 2013)

SSHRC≡CRSH   The Social Sciences and Humanities Research Council (SSHRC)

CIHR IRSC   The Canadian Institutes of Health Research (CIHR)

NSERC CRSNG   The Natural Sciences and Engineering Research Council (NSERC)

INNOVATION.CA   The Canada Foundation for Innovation (CFI)

Feedback: Email responses should be sent to digitalscholarship@sshrc-crsh.gc.ca.

Government   Gouvernement
of Canada   du Canada

Table of Contents

Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada

Consultation Document
(16 October 2013)

Executive Summary

Through various cooperative investments and ventures involving post-secondary institutions, not for profit organizations and virtually all levels of government, Canada has put in place many of the elements of a well-functioning digital infrastructure ecosystem for research and innovation. While this has provided the foundational platform for a large segment of the research community, it is equally clear that the potential of data-intensive research is progressively and rapidly outstripping our ability to manage and to grow the digital ecosystem required to meet 21st century needs.

Canada's research funding agencies, the Social Sciences and Humanities Research Council (SSHRC), the Natural Sciences and Engineering Research Council (NSERC), the Canadian Institutes of Health Research (CIHR) and the Canada Foundation for Innovation (CFI)—collectively the TC3+, in collaboration with Genome Canada—have joined forces to address this challenge. Through the attached consultation document, they are proposing a collective realignment of agency funding policies regarding management of data obtained through projects undertaken with agency funds. Specifically, the following initiatives are proposed:

1. **Establishing a Culture of Stewardship** – The onset of the "data deluge" threatens to outpace the evolution of a culture of data stewardship within the research community. To address this, research funding agencies, research institutions and professional scientific associations—with reference to existing data and best practices globally—should cooperate in the development of clear policies and guidelines to bring attention to this important aspect of research culture and promote the development of appropriate data management systems and capabilities. To this end, in consultation with stakeholders and the research community, we propose that the TC3+ define the core elements of an agency-based and focused data stewardship plan.

2. **Coordination of Stakeholder Engagement** – The Canadian research environment is characterized by a high degree of commitment to collaboration and cooperation, with individuals and organizations combining forces to address long-term planning and initiate "bottom-up" actions to shape the evolution of the digital landscape. These important actions, which bring a richness of community engagement, would benefit from the development of a coordinating mechanism to ensure enhanced alignment, helping to optimize returns on the time invested by all parties. To encourage maximum contribution from key players, the TC3+ would work with other organizations and working groups to ensure ongoing consultation and coordination with all stakeholders, including the provinces, in the development of Canada's national digital infrastructure ecosystem for research.

3. **Developing Capacity and Future Funding Parameters** – To help create a forward-looking digital research environment, the parameters for the funding of coordinated national-scale digital infrastructure should be re-examined. In particular, the balance of roles and responsibilities among national, provincial and institutional stakeholders should be reassessed to ensure both effective support and efficiency. For its part, the TC3+ proposes broad collaboration in the development of a coordinated plan to encourage the establishment and

sustained operation of a number of world-class centres specializing in data management and supporting interrelated functions including but not limited to administration, operations, policy and access; enhanced networks and infrastructure; skills development; and graduate and researcher training.

The goals of such actions are threefold: 1) to encourage the collection and—in particular—sharing of data among researchers, in keeping with the Government of Canada's Open Government initiative (www.data.gc.ca); 2) to encourage further development of an environment for digital scholarship in Canada that provides for more effective coordination among the various stakeholder organizations; and 3) to help ensure that existing and future investments in digital infrastructure and training are maximized to the benefit of Canadians.

The document is now being forwarded to stakeholder and research communities as part of a broad consultation process informing future development of the proposed policy framework and associated next steps.

Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada

Introduction

This document has been prepared jointly by the federal Tri-Council Agencies (the Social Sciences and Humanities Research Council, the Natural Sciences and Engineering Research Council and the Canadian Institutes for Health Research) and the Canada Foundation for Innovation (collectively referred to as the TC3+), in collaboration with Genome Canada, and proposes changes to their funding policy frameworks that promote excellence in data management practices, thereby advancing digital scholarship and Canada's digital infrastructure ecosystem to the benefit of Canadians. The specific proposals contained herein (p. 11), respond to widespread changes associated with the rapid development of digital scholarship globally and to their resulting impact on the research enterprise in this country. As part of a national consultation process designed to inform their ultimate content, they are hereby presented to the Canadian digital scholarship stakeholder and research communities for broad discussion and input.

The Changing Environment for Digital Scholarship

Canada currently finds itself in the midst of an information and communications revolution of transformative social, economic and cultural impact, involving deep conceptual changes that have been increasingly enabled, accelerated and influenced by dynamic new technologies. Significantly, this revolution reflects both the technologies themselves and the massive amounts of information—data— that these tools now capture and process, whether related to consumer behaviour, astronomical observation or human health. Indeed, it is this vast increase in the flow of data, from a variety of sources, which has come to represent a significant source of innovation and "disruption" in and of itself.

> "We all experience it: a rising tide of information, sweeping across our professions, our families, our globe. We create it, transmit it, store it, receive it, consume it—and then, often, reprocess it to start the cycle all over again. It gives us power unprecedented in human history to understand and control our world. But, equally, it challenges our institutions, upsets our work habits and imposes unpredictable stresses upon our lives and societies."
>
> The European Commission, 2010[1]

As this transformation moves into an accelerated and expanded phase, the focus of data analysis is rapidly shifting to embrace not simply technical development, but also new ways of thinking about social, economic and cultural expression and behaviour. Indeed, innovative information and communications technologies are enabling the transformation of the fabric of society itself, as data becomes the new currency for research, education, government and commerce. As such, data are rapidly becoming a torrent flowing into every area of the global economy, society and culture.[2] Recent studies suggest that, fueled by enormous flows of data, we are in fact at a potential tipping point of a tremendous wave of exploration, innovation, productivity and growth, as individuals, post-secondary

---

[1] European Commission. *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. The European Commission. October 2010.

[2] McKinsey & Company. *Big Data: The next frontier for innovation, competition, and productivity*. May 2011.

institutions, companies, governments and organizations of all types begin to exploit its potential. Further, the ability to capture, analyze and share vast amounts of data is fundamentally enabling new research approaches and practices, such as bioinformatics and computational biology, creating what Tony Hey from Microsoft Research has called the fourth paradigm of scientific exploration.[3] This new paradigm is also redefining the interrelationships between the traditional categories of post-secondary teaching, research and service. Indeed, data are now, at one and the same time, research data, learning data and innovation data. This shift also brings with it deep concerns related to issues such as training, analytic capacity, accessibility, ethics, intellectual property and privacy.

As with all aspects of the information and communications revolution, the extent to which technological development and data are successfully exploited to fuel economic activity and enhance our society depends upon human decision-making. The way that we, as a country, choose to manage our data will directly impact our ability to undertake leading-edge research and development in the future and to benefit from that activity. Data and information do not respect borders, either on campus or beyond; managing data is about much more than supporting research excellence. Digital data are the raw materials of the knowledge economy, and are becoming increasingly important for all areas of society, including industry. In their 2011 report on Big Data, McKinsey Global Institute pointed out that, "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". The same may be said of the capacity to capture, manage and preserve it, or the requisite training of personnel who can operate effectively in this milieu.

Canada now stands in direct competition with a host of other countries, including the United States, European Union countries, Australia and other technologically advanced countries, in the race to develop an effective strategy for harnessing the digital wave. A coordinated approach to policy formulation; personnel training; development of infrastructure and analytical tools; and managing, conserving and providing access to research data would help ensure that Canadians derive greater and more long-term benefit, both socially and economically, from the extensive public investments that are made in research. Such an approach would represent a logical extension of the Government of Canada's Open Government initiative (www.data.gc.ca), which already aims to make government-generated data widely available.

> "Although only a fraction of such data is intended or destined for publication in journals, the potential utility of these separate, disconnected data stores is profound. Any given tidbit of data produced by one researcher might supply a missing puzzle piece to another—even one involved in seemingly unrelated work. The ability to harness such data and apply it in new ways and new directions holds the promise of substantially accelerating research and innovation."
>
> Report from the Thomson Reuters Industry Forum, 2013[4]

A   central, emerging public policy question is then how to put data-intensive, computationally powerful modes of discovery to use in both the public and private sectors to support jobs, productivity and economic growth, stimulate innovation, and find new ways to address global challenges such as climate change,

---

[3] Tony Hey. *The Fourth Paradigm: Data Intensive Scientific Exploration and Discovery*, Microsoft.
[4] Thomson Reuters. *Unlocking the Value of Research Data*. Report. July 2013.

health-care delivery, poverty reduction, energy security or medical diagnostics and disease prevention. At the same time, we must find ways of addressing the complex ethical, legal and social questions raised by new forms of data collection, data mining and data sharing. This is particularly important as smart phones, social networking and massive consumer and health service user databases allow us to track and analyze human behaviour at the level of individuals.[5]

Although the challenges created by what Alex Pentland from MIT's Media Lab has called a "data-driven society" appear daunting, Canada is well placed to both take full advantage of the social and economic potential of the data deluge, and discover new ways of ensuring confidentiality, privacy and the protection of public interests. For example, recent reports have positioned information and communications technology (ICT) as one of the largest public/private research enterprises in the country in terms of output, and have indicated that the field of ICT accounts for 44 percent of Canada's patents. A sustainable world-class advanced digital infrastructure ecosystem is—without question—one critical driving factor for economic success in this sector and potentially a range of others.

Canada currently possesses the necessary components of digital infrastructure needed to support the creation of such an ecosystem, including high performance computing, high performance networking, data repositories, highly qualified personnel, and the related tools and services. However, the organizations and service providers have lacked to date a single coherent and cohesive vision. Moving to establish greater coordination and coherency will then have no other effect than to create a more efficient and effective support system for data-driven research and technology development. Further, it will reduce our productivity gap with respect to our competitors in the global marketplace, accelerate the pace of innovation, and enable economic growth and wealth creation through all sectors of the economy.

System Needs for Effective Digital Infrastructure

Today, one of the principal challenges associated with the new digital environment is related to both infrastructure and management of volumes of data: how data are coded for preservation and access, how they flow, how they are stored, made accessible and translated into knowledge. In many cases, we are no longer data limited; rather we are "insight" limited. Indeed, the 2013 e-science conference in São Paulo emphasized that the current key digital challenge is now "Turning Data into Insight." For this purpose, we need an advanced digital infrastructure ecosystem that supports the seamless access, use, re-use and integrity of data, and to focus on establishing and operating the processes required to collect, manage, analyze, interpret, share and archive big data. This ecosystem must integrate means for researchers from all sectors to utilize the technology effectively since the human infrastructure is as important as the technological. There also must be coherence, coordination and alignment across the diverse elements of the digital infrastructure.

---

[5] Steve Lohr. *The Promise and the Peril of the 'Data Driven Society'*. The New York Times. February 2013.

There are various ways to characterize an advanced, robust digital infrastructure ecosystem. The following principles help provide a reasonable starting point (adapted from the report of the 2012 Digital Infrastructure Summit):

- **Integration** – all essential elements are integrated in a national system that avoids unnecessary duplication, fragmentation and overlap of service; similarly avoiding unnecessary competition.
- **Inclusivity** – dealing with all federal, provincial/territorial, regional, institutional stakeholders in all sectors and research communities.
- **Sustainability** – designed and supported in a way that enables evolution and adaptability, agility and responsiveness to users, stakeholders and technology evolution.
- **Comprehensive** – provides a full spectrum of service dealing with all parts of the evolving digital infrastructure ecosystem (e.g., networks, computational facilities, tools and services, data management, the framework conditions, people).
- **Accessible** – maximum ease of access, regardless of location, discipline or level of expertise; a "one-stop-shop" approach is desirable from a user's perspective as opposed to having to negotiate with many service providers.
- **Valued** – stakeholders support the national system; user services are prioritized based on the value stakeholders place on them, while being responsive to opportunities in the evolving ecosystem.
- **Governed** – the national system has effective oversight, stakeholder engagement and alignment of its component parts.
- **Agil** – responsive to user needs, focused on rapid and efficient service delivery.
- **Ethical** – supports and promotes the ethical use of data and enables freedom of inquiry.

The State of Canada's Digital Infrastructure

Through various cooperative initiatives and investments among the federal government, the provinces/territories, institutions and—in many cases—industry, Canada has put in place many of the elements of a well-functioning advanced digital infrastructure for research and innovation. This in turn provides the foundational platform for a large component of the research community. We can, in fact, be proud of the world-class high bandwidth, low latency network across Canada (managed by CANARIE and the Optical Regional Advanced Networks—ORANs) and the provision of pan-Canadian high-performance computing facilities and services by Compute Canada. And there is increasing attention being paid to data management and personnel training by funders, institutions, researchers and librarians. As yet, however, the potential of data-intensive research has outstripped our ability to manage and to grow the broader digital infrastructure ecosystem required to meet 21st century needs.

At the moment, there are a large number of players and stakeholders inhabiting the digital landscape, all of who need to be engaged in the ecosystem development process. These include:

The Funders
The federal Tri-Council funding agencies play a large role in the structuring and funding of digital infrastructure and advanced training in Canada. The division of responsibility has evolved since the formation of CFI in 1997 and Genome Canada in 2000. Over the past 15 years, CFI in particular has gradually taken on the role of prime funder for computational resources that are above and beyond those accessible from a project grant. Most importantly, it has also taken a lead in rationalizing high-performance computing services.

Provinces (and potentially territories) are also key funders, in some cases through their funding agencies and in others through line ministries—as is manifest through their operating support for post-secondary institutions, e-enabled health-care systems (including hospitals) and the ORANs. Matching funding from institutions and the private sector constitutes another source of support for infrastructure.

The Service Providers
The major service providers in Canada are CANARIE, the ORANs and Compute Canada. The roles of CANARIE in providing a broadband low latency network and Compute Canada in developing computational resources are well understood. CANARIE has recently taken over management of the Canadian Access Federation—a trusted access management environment (single-identity access) for the Canadian research and higher education communities.

There are also content service providers including: 1) the Canadian Research Knowledge Network (CRKN); 2) the National Research Council of Canada, which manages DataCite (Canada's data registration service); and 3) CASRAI, a standards organization that deals with research administration data that is expected to include standards for metadata.

Additionally, there exist national domain-specific specialized research data management infrastructures: 1) the Canadian Research Data Centre Network (CRDCN), which provides services, some curation, and protected access to Statistics Canada data and is now expanding to include other confidential datasets; 2) the Canadian Polar Data Network (CPDN), which is an outgrowth of the very successful data management activity that was developed to support the International Polar Year (IPY); and 3) the Canadian Astronomical Data Centre (CADC), which is an internationally connected trusted repository for astronomy data collected through projects involving Canadian researchers.

Finally there are project-level research data-management infrastructures such as Ocean Networks Canada and the Canadian Longitudinal Study on Aging.

Contributors/Beneficiaries
Clearly academic research institutions and their research communities are both contributors to and key beneficiaries of a robust advanced digital infrastructure. Those institutions are variously represented in their interest in this area by:
- Presidents, Vice-Presidents Research, and other senior administrators at post-secondary institutions and hospitals;
- IT Units, Chief Information Officers (CIOs) and their national organization, the Canadian University Council of CIOs (CUCCIO); and
- university libraries, Chief Librarians and their national organization, the Canadian Association of Research Libraries (CARL).

Regional innovation systems are also beneficiaries through private sector use of CANARIE, and the local/regional ORANS as well as access to the research and talent supported through the Tri-Council, CFI and Genome Canada.

Collaborative Groups
Two groups have emerged from stakeholder concerns regarding the sustainability and future development of Canada's advanced digital infrastructure ecosystem. The first to emerge was the Research Data Strategy Working Group (RDSWG), which morphed into Research Data Canada in 2012, a

forum for stakeholders to work together to enhance research data stewardship. The second was the Leadership Council for Digital Infrastructure, which is focusing on development of a national strategy to renew and strengthen Canada's advanced digital infrastructure ecosystem with data management, sustainable funding and integrated planning identified as key areas to be addressed.

Toward Policy Development and the Establishment of Effective Framework Conditions

Given the plethora of stakeholders and the complexity of the system, there is considerable room currently for enhanced, pan-national strategic thinking and action on issues that are central to the future health of Canada's advanced digital infrastructure ecosystem. Indeed, through a relatively limited number of initiatives, an environment could be fostered that can better guide and ensure the continued development of forefront research and scholarship that would be of use to both the public private sectors.

In response, what the TC3+ is proposing in this document is not yet another agency or organization to oversee the broad array of assets and services now in the digital infrastructure domain. Rather, we are suggesting a realignment of agency funding policies, primarily by fostering greater clarity and overall coordination with respect to role of the federal Tri-Council funding agencies and the CFI in their support of digital scholarship across all research fields. This in turn, we believe, will provide an initial but critical step toward the creation of an environment for digital scholarship and training in Canada that provides clarity and effective coordination for the various stakeholder organizations—including other funders, service providers and research performers—to ensure that their objectives and efforts contribute optimally to specific definable goals and objectives related to Canada's digital infrastructure ecosystem.

In pursuit of this objective and in order to enhance traction toward a robust and sustainable digital infrastructure ecosystem for Canada, the following set of proposed actions is presented by the Tri-Council and the CFI for discussion and input from stakeholders and the broader research community.

1. **Establishing a Culture of Stewardship** – The onset of the "data deluge" threatens to out-pace the evolution of a culture of data stewardship within the research community. To address this situation, research funding agencies, research institutions and professional scientific associations, with reference to existing data and best practices globally, should cooperate in the development of clear policies and guidelines to give profile to this important aspect of research culture and promote the development of appropriate data management systems and capabilities. To this end, we propose that the TC3+ define the core elements of an agency-based and focused data stewardship plan. The elements of this plan could potentially include:

    a. a requirement that all grant applications include specific data management plans including identified costs of data collection/analysis and preservation of results and associated datasets;
    b. definition of those specific elements of data plans that will be considered by reviewers in the assessment of funding applications;
    c. guidelines indicating which data must be preserved and in what formats;
    d. consolidated open access policies and guidelines (in concert with work already initiated by the TC3+);
    e. guidelines for researchers in selecting suitable data repositories;

f. recognition of data repositories across Canada that meet global standards for such facilities; and

g. guidelines for ensuring informed consent for data use and protection of privacy and confidentiality.

2. **Coordination of Stakeholder Engagement** – The Canadian research environment is characterized by a high degree of commitment to collaboration and cooperation, with individuals and organizations combining forces to address long-term planning and initiate "bottom-up" actions to shape the evolution of the digital landscape. These important actions, which bring a richness of community engagement, would benefit from development of a coordinating mechanism to ensure enhanced alignment to optimize returns on the investment of time by all parties. To help ensure the maximum contribution of key players, the TC3+ would work with other organizations and working groups to ensure ongoing consultation and coordination with all stakeholders, including the provinces, in the development of Canada's national digital infrastructure for research.

3. **Developing Capacity and Future Funding Parameters** – To help create a forward-looking digital research environment, the parameters for the funding of coordinated national scale infrastructure should be re-examined. In particular, the balance of roles and responsibilities among national, provincial and institutional stakeholders should be reassessed to ensure both effective support and efficiency. For its part, TC3+ would collaborate in the development of a coordinated plan to encourage the establishment of new and/or the enhancement and sustained operation of existing world-class centres specializing in data management, supporting:

   a. best practices in administration, operations, policy and access;
   b. enhanced networks and infrastructure, as well as support for enhanced research on; for example:
      i. identifying further gaps and proposed solutions in the development of Canada's digital scholarship infrastructure, technology and services;
      ii. identifying best practices globally in data management;
      iii. aligning and incorporating inclusion of other prospective partners in digital scholarship and data management, with a specific focus on private sector participation; and
      iv. establishing the best means for international collaboration in data management;
   c. skills development and graduate and researcher training.

These steps put forward by TC3+ are proposed to accelerate the overall move toward a coordinated and effective approach in which a shared, common understanding informs action by the various contributors to the digital landscape in Canada. This in turn will produce specific concrete results, including increases in both the volume of research data available for public and private sector access and in the development of talent that will help make best use of the ecosystem. Overall, our hope is that by arriving at an agreed-upon articulation of the key features of a robust and sustainable digital infrastructure ecosystem and by explicitly coordinating the diverse and multiple steps forward needed to move toward this ecosystem, Canada can quickly advance as a global leader in building prosperous, resilient and just societies in the Digital Age.

Appendix 1: The Components of the Digital Infrastructure Ecosystem

- **The framework conditions** – the policies and legal framework within which digital research is undertaken; the means of coordination and alignment of various components of the digital research environment; the suitability of funding systems and reward systems for pursuit of e-research; the capacity of Canada to deal with other international players in digital research.
- **Expertise and skills** (as an input) – the sufficiency and quality of skilled personnel, both generic and domain specific, for effective use of the components of the e-infrastructure.
- **Tools and services** – the software, applications and human support services that enable researchers to derive value from their data and to optimize the use of the research instruments and systems.
- **Research data and research data management infrastructure(RDMI)** – both data as infrastructure and systems for managing data—the collection, structuring, standardizing, archiving, curating and sharing of with system characteristics of flexibility, security, accessibility, interoperability, affordability, open access and high performance.
- **Computational facilities, tools and services** – hardware and associated software resources that enable both compute-intensive and data-intensive research, as well as the services and tools that enable value to be derived from the facilities. This includes both Cloud and Grid computing.
- **Networks and tools and services** – means of connecting researchers to data sources and transporting data among different locations.
- **Collaboration infrastructure and tools** – means of connecting researchers to researchers and partners in multi-sectoral research initiatives that are geographically dispersed and/or are utilizing common datasets and tools.

Appendix 2: Stakeholders and Landscape Map

Brief descriptions follow of the various organizations and working groups active in promoting and supporting digital scholarship and the cyber-infrastructure underpinning in Canada.

Research Data Canada (RDC)
- **What** – A stakeholder-driven and supported national body dedicated to advancing the vision for research data in Canada.
- **Who** – A variety of stakeholder organizations, all with an interest and role to play in ensuring that the infrastructure, processes and support are in place to realize the vision for research data in Canada. This includes CUCCIO, Compute Canada, CANARIE, CARL, CFI, CIHR, NSERC, SSHRC, CASRAI, NRC, TB, IPY, LAC and CODATA.
- **Mandate** – To develop strategy, facilitate communication and partnerships among data initiatives, promote education and training in data skills, measure progress in implementing the vision, bring attention to gaps, and act as a single point of contact for Canada in international data initiatives (self-generated mandate).
- **Functions** – Its activities focus on five areas: policies, infrastructure, standards and interoperability, education and training, and international liaison. RDC does not plan to own or operate infrastructure.
- **Intersects/interfaces** – May be somewhat different from the Digital Infrastructure Leadership Council in that it focuses on issues related to research data and data lifecycle management. However, there does appear to be some overlap between the two in focus and participation.
- **Recent actions** – The inaugural RDC meeting was held in January 2013. RDC's priorities for 2013 are to:
    - **launch Research Data Canada** as the multi-stakeholder/volunteer-driven organization that will drive efforts forward to ensure that the full value of Canada's research data is realized;
    - **encourage broad membership** for Research Data Canada to reflect fully the diversity of stakeholders with an interest in research data;
    - **advance the work of the RDC Committees**: Infrastructure, Education and Training, Policy, Standards and Interoperability, and International Liaison;
    - **host Webinar series** on data management – Canadian and international speakers, range of topics, audiences;
    - **co-sponsor data stream in CASRAI Big Data Reconnect conference** October 2013 and coordinate a pre-conference Data Centres workshop;
    - **continue international liaison** with Research Data Alliance, Global Research Data Infrastructure, DataCite Federation and other initiatives;
    - **establish a national advisory council** of senior representatives from industry, the academy, government research labs, funding agencies and policymakers to provide counsel to Research Data Canada; and
    - **initiate a national online consultation** process to take the results of the 2011 Canadian Research Data Summit to a broader set of stakeholders across the country.

- **History:**
    - Took over from the Research Data Strategy Working Group (RDSWG) in 2012; originally formed to survey and identify the challenges and issues surrounding access to, and preservation of, data arising from Canadian research;

- The RDSWG organized the September *2011 Canadian Research Data Summit: Mapping the Data Landscape*;
- The final report of the Summit proposed a National Strategy for Research Data in Canada, including a vision statement, high-level goals and a framework for action with broad timelines and distribution of tasks across major stakeholder communities; and
- Also conducted a gap analysis for Canada: http://publications.gc.ca/collections/collection_2009/cnrc-nrc/NR16-123-2008E.pdf
- The RDSWG transferred its activities to Research Data Canada in December 2012.

Leadership Council for Digital Infrastructure

- **What** – A stakeholder-driven and supported national initiative dedicated to developing a national strategy to renew and strengthen our current advanced digital infrastructure for research, innovation and education in Canada.
- **Who** – A cross-sector group co-chaired by Jay Black, SFU and CUCCIO, and Steven Liss, Vice Principal Research, Queen's University, with membership including: CRKN, CUCCIO, Compute Canada, CANARIE, CARL, CFI, NRC, Industry Canada, CIHR, NSERC, SSHRC and CASRAI.
- **Mandate** – To build on the work accomplished at the Summit, develop plans for first initiatives, and put in place the mechanism to ensure continued engagement of the many stakeholders.
- **Functions and priorities** – Conduct a gap analysis, develop a roadmap to address the gaps and convene a follow-up to the 2012 Summit.
- **Intersects/interfaces** – Different from Research Data Canada in that its focus is to provide an overarching view and the required mechanism(s) to support an integrated and sustainable approach to Canada's advanced digital infrastructure eco-system.
- **Recent actions** – Were a result of the CUCCIO-hosted Digital Infrastructure Summit 2012 in Saskatoon.

Canadian Association of Research Libraries (CARL)

- **What** – The leadership organization for the Canadian research library community.
- **Who** – Members are Canada's 31 large research libraries.
- **Mandate** – Provides leadership on behalf of Canada's research libraries and enhances their capacity to advance research and higher education. It promotes effective and sustainable scholarly communication, and public policy that enables broad access to scholarly information.
- **Functions and priorities** – Its *2013-2016 Strategic Directions* include the following points:
    - facilitate collaborations to share and preserve Canada's research collections;
    - coordinate research data management initiatives; and
    - promote open access and new forms of scholarly communication.
- **Recent actions related to digital scholarship**:
    - CARL led the development of a concept of a distributed network of data repositories, which was to inform a 2012 proposal for CFI funding; ultimately, a formal proposal was not submitted to CFI as the LEF-NIF was not an ideal CFI program for it.
    - Along with CRKN, CARL published in October 2012 a report emphasizing the potential role(s) for academic libraries in implementing an open access policy on research publications.
    - In January 2013, CARL held a four-day "Introduction to Research Data Management" course for librarians; about 60 individuals enrolled from 30 universities; these participants have continued to confer as an online "community of practice."

- CARL has worked to facilitate the development of open access digital repositories at (by now) all member libraries; most CARL member libraries are developing local research data management services for researchers.
- CARL has advocated for federal government support for national research data management infrastructure; it has developed web content, conference presentations, workshops and articles on both open access and research data management.
- CARL is working with CASRAI and Research Data Canada to produce CASRAI's 2013 Big Data conference.
- CARL has created a Data Management Subcommittee headed by its Vice-President/President-Elect (Martha Whitehead, Queen's University) that is proposing the formation of a national collaborative network of local/regional/other initiatives for collecting, preserving and providing access to research data produced in Canada.

- **Intersects/interfaces** – Both CRKN and Canadiana.org (Canada's premier organization for the digitization of Canadian historical documentation and the exposure of Canadian digital documentary collections) had their origins as CARL initiatives; CARL collaborates closely with them. CARL is a supporting member of Research Data Canada and the Leadership Council for Digital Infrastructure. CARL is the major Canadian association member of SPARC (the Scholarly Publishing and Academic Resources Coalition) and COAR (the Coalition of Open Access Repositories), both of which promote and support open access and repositories internationally. Locally, CARL member library directors work in consultation with CUCCIO member CIOs in the context of various digital services.

Ontario Council of University Libraries (OCUL)
- **What** – A library consortium
- **Who** – Ontario's 21 university libraries
- **Mandate** – To enhance information services in Ontario and beyond through collective purchasing and shared digital information infrastructure, collaborative planning, advocacy, assessment, research, partnerships, communications and professional development.
- **Functions** – Provide access to a diversity of learning and research materials, and ensure their preservation through sustainable and responsible stewardship; lead in the development of partnerships to expand Canada's digital research infrastructure; operate Scholars Portal; encourage the advancement of access to electronic data resources including those provided under the Data Liberation Initiative (DLI); expand access to maps, geospatial data and other cartographically related resources, both print and digital.
- **Recent activities** – Scholars Portal has received certification as the first Trustworthy Digital Repository in Canada. This certification, the only generally recognized certification for digital archives, was issued by the Center for Research Libraries (CRL).

Canadian University Council of Chief Information Officers (CUCCIO)
- **What** – CUCCIO is a non-profit, member-funded association of Canada's higher education information technology leaders, working together to help Canadian universities excel through the innovative and effective use of IT.
- **Who** – Composed of the chief information officers (CIO) from more than 50 universities Canada-wide.
- **Strategic priorities/mandate** (from Strategic Plan):
  - foster best practices in information technology management in Canadian universities;
  - identify, incubate and sponsor collaborative sector-wide services;

- — develop and deliver programs and services to support the professional development of IT staff; and
  - — develop and maintain relationships with governments, government agencies, corporations and other groups of interest to higher education in order to advance the shared interests of Canadian universities.
- **Recent actions related to digital scholarship:**
  - — Convened the *Digital Infrastructure Summit 2012* to:
    - ▪ establish a vision for a comprehensive, integrated and sustainable digital infrastructure in support of research, education and innovation in Canada;
    - ▪ develop a specific action plan, with milestones; and
    - ▪ secure commitments from stakeholders with regards to realizing the plan.
  - — Emanating from the Summit, it created a cross-sector group, the *Leadership Council for Digital Infrastructure*, to build on the work accomplished at the Summit (above).
  - — Plans are underway for a follow-up to Summit 2012.
- **Intersects/interfaces** – The work of the Council will intersect with the strategic and operational plans of Compute Canada, CANARIE and Research Data Canada from the perspective of the digital infrastructure overall and to facilitate as possible integration and coordination of the various components.

Compute Canada

- **What** – An incorporated NFP organization that provides Canada's national platform of High Performance Computing (HPC) resources. A new President and Board are in place effective late 2012.
- **Who –** Members are 29 Canadian universities. Membership is available to any university or college in Canada that has one or more researchers using an advanced computing system, through access provided by Compute Canada. Compute Canada assigns services on a pan-Canadian basis, using regional nodes:
  - — Compute West – WestGrid
  - — Compute Ontario – HPCL, SciNet, SHARCNet
  - — Calcul Quebec – formerly RQCHP and CLUMEQ
  - — Compute Atlantic – ACENet
- **Functions –** Compute Canada delivers its services through HPC systems managed by regional consortia at different locations across Canada and utilizing the CANARIE broadband network.
- **Mandate** – To promote and support the shared use of advanced computing resources designed to keep Canada competitive in research and innovation. It is not, however, a policy-making body or a standards organization.
- **Assets** – The Compute Canada platform includes computing capability, online and long-term storage, connection to the CANARIE network, and user support services. It is primarily oriented toward larger computation systems used in simulation and computational intensive research.
- **Funding** – Capital assets funded in part by CFI. Institutions also a player. Also in receipt of significant support for operating and maintenance from the CFI MSI Fund.
- **Recent and current actions:**
  - — Recent unsuccessful proposal to CFI (LEF-NIF competition) to address the needs of the medical and SS&H communities for:
    - ▪ "big data" systems; i.e., systems with high "storage to processing" ratios (far beyond the current Compute Canada offering) and high throughput; and
    - ▪ secure data processing environments (e.g., for confidential data).

- — Issue at the time was more capacity than direction (which was deemed good).
- — Development of a strategic plan for the organization through 2013.
- • **Intersects/interfaces –** Provinces, CANARIE, RDC, Leadership Council, private sector.

CANARIE

- • **What** – CANARIE supports research and education through the delivery of advanced digital infrastructure. Manages and evolves one of the world's largest and fastest research and education networks, in partnership with provincial and territorial networks (ORANs).
- • **Who** – A not-for-profit corporation funded primarily through the federal government, with additional funding from membership revenues and fees for services.
- • **Mandate** – to design and deliver digital infrastructure, and drive its adoption for Canada's research, education and innovation.
- • **Functions** – manages an ultra-high-speed national backbone network that connects provincial and territorial networks to each other and to 100 international networks. The provincial and territorial networks connect directly to universities, research centers, government labs, hospitals and other scientific facilities within their jurisdictions, and to CANARIE's national backbone and global research and education networks. The partnership of CANARIE and the ORANs enables researchers, educators and innovators to move, share and analyze data and access specialized tools and resources. CANARIE also supports the development of software to support research collaboration and access to widely distributed data and tools. CANARIE spurs research and innovation in the private sector by offering small and medium-sized businesses access to a cloud-based testbed to accelerate product development timelines and reduce costs.
- • **Objectives** for the period April 1, 2012 to March 31, 2015 are:
  - — **Network Operations –** Operation and evolution of the CANARIE network as essential research infrastructure; extending the "owned" portions of the network to provide greater flexibility and lower costs for delivering greater bandwidth, including network-based services, which currently include:
    - ▪ [Canadian Access Federation (CAF)](#);
    - ▪ [Content Delivery Service (CDS)](#); and
    - ▪ managing the Network Alliance Infrastructure and Network Alliance Development programs to strengthen the pan-Canadian network and enhance the visibility of this essential digital infrastructure.
  - — **Technology Innovation** – Develop, demonstrate and implement next-generation technologies to advance the CANARIE network as a leading-edge research network— including new software tools, comprising a toolkit of reusable services. Two programs support this objective:
    - ▪ Research Platform Interfaces (RPI) – Leverages services from the previous Network-Enabled Platforms (NEP) program by creating a collection of platform services (RPIs) from existing NEPs to be used by multiple research platforms; and
    - ▪ Network-Enabled Platforms (NEP) – The development of sophisticated software platforms that enable researchers to easily collaborate and access research data and tools.
  - — **Private Sector Innovation** – Leveraging the CANARIE network to assist firms operating in Canada, and Canadian universities, to advance innovation and commercialization of products and services to bolster Canada's technology innovation capabilities. CANARIE's DAIR (Digital Accelerator for Innovation and Research) program offers a cloud-computing testbed for small and medium-sized enterprises to accelerate product development, reduce costs and realize the scale and agility benefits of cloud technologies.

ORANs
- **What** – Optical Regional Advanced Networks (ORANs)
- **Who** – 12 regional networks:
    — East – ACORN-NL, ACORN-NS, New Brunswick Advanced Network, Prince Edward Island Advanced Network;
    — Quebec – RISQ;
    — Ontario – ORION;
    — West – BCNet, Cybera, SRnet, MRnet; and
    — Territories – Aurora College, Yukon College.
- **Mandate** – To support the operation and development of advanced networks and services at the regional level in support of research and innovation.
- **Intersects/interfaces** – The ORANs provide connectivity of the regional high-speed network to the national backbone provided by CANARIE.

Canadian Access Federation
- **What** – A trusted access management environment (single-identity access) for Canadian research and higher education communities.
- **Who** – Developed by CUCCIO with operational responsibility transferred to CANARIE in 2012.
- **Mandate** – To make sharing protected resources easier, safer and more scalable by:
    — enabling staff, students and faculty to access wireless networks and web-based resources using their home organization credentials when they are visiting other organizations;
    — allowing participants to participate in a cost-effective, privacy-preserving approach to access management;
    — helping to ensure the privacy of personal information by eliminating the need for researchers, students and educators to maintain multiple, password-protected accounts; and
    — enabling organizations to better manage access to their resources based on a user's status and privileges as presented by the user's home organization.

Canadian Research Knowledge Network (CRKN)
- *What* – An organization of Canadian universities dedicated to expanding digital content for the academic research enterprise in Canada.
- **Who** – An incorporated NFP organization that is a partnership of 75 Canadian universities. University libraries are the drivers of CRKN's initiatives, and play a primary role in leveraging expertise and resources for the benefit of Canada's scholarly research community.
- **Mandate** – To undertake large-scale content acquisition and licensing initiatives in order to build knowledge infrastructure and research capacity in Canada's universities and to provide equitable and cost-effective access to scholarly content for universities nationwide.
- **Recent activities** – CRKN's current draft strategic plan identifies a role for the organization in supporting and coordinating vertical integration of all types of research data and digital scholarship nationally.
- **Intersects/interfaces** – Involves all of the AUCC libraries including the members of CARL plus approximately 40 others. CRKN has worked with CUCCIO to develop the Canadian Access Federation, and collaborates with regional academic library consortia including COPPUL, OCUL, CREPUQ and CAUL. CRKN represents Canada on the SCOAP3 Open Access initiative.

NRC – CISTI (Canadian Institute for Scientific and Technical Information)
- **What** – Canada's national science library.
- **Who** – CISTI is a division of NRC.
- **Mandate** – Undergoing a major transformation from science journal publisher/library to an electronic library service provider.
- **Recent activities***:*
  — CISTI is currently building a mirror site for the NIH PubMed Central. Funded by CIHR, this site will provide Canadian researchers with access to American publications in PubMed Central and provide a portal for the deposit of Canadian journal articles. The portal will allow CIHR to eventually develop a fully mandatory open access policy.
- **Intersects/interfaces** – CISTI has also acted as host and administrator for various data and open access working groups, including the new body, Research Data Canada—activities that have linked it with the funding agencies, CARL, CASRAI, the CRDCN, etc.

CASRAI
- **What** – A not-for-profit standards development organization focusing on research administration data. The Board represents the diversity of stakeholders.
- **Who** – A community of research organizations (funders—federal and provincial, institutions, implementers) collaborating to evolve the standard dictionary of research terminology and to advance the standard platform for research interoperability. Representatives from participating organizations sit on a number of committees, review circles and advisory councils. There are international mirrors of CASRAI Canada developing though the leadership of CASRAI.
- **Mandate** – To provide a forum and the mechanisms required to standardize the data that researchers, their institutions and their funders must produce, store, exchange and process throughout the life-cycle of research activity.
- **Priorities** – To advance those semantic standards that will facilitate effective operation in the digital business environment.
- **Intersects/interfaces** – Participates in the Leadership Council and the RDC.

Canadian Research Data Centre Network (CRDCN)
- **What –** The Network acts as a pan-Canadian forum and structure to give Canada's research community access to social and population health statistics and help provide evidence for effective public policy and planning. CRDCN oversees a pan-Canadian array of Research Data Centres (RDCs). An RDC is a university-based laboratory, staffed by a Statistics Canada Analyst, that offers researchers on-site services for:
  — **Secure access to confidential micro-data** – Statistics Canada census and surveys, plus a growing range of administrative data; prospects of datasets from other federal departments.
  — **What they need to analyze the data** – Fully-equipped workstations, statistical software and technical support
- **Who –** 45 academic institutions and Statistics Canada form the core membership of the Network.
- **Mandate:**
  — **To improve data access** by giving researchers across the country access, free-of-charge, to detailed micro-data from an increasing range of survey, census and administrative data.
  — **To expand the pool of skilled quantitative researchers** in Canada and train the next generation of researchers.

- — **To make research count** by improving communication between social scientists and the potential users of the knowledge they create.
- **Intersects/interfaces** – Statistics Canada, CANARIE, the ORANs, NRC-CISTI, various social and health policy federal departments and agencies.

Canadian Polar Data Network
- **What** – A Canadian network and standards-based organization that grew out of the data centre for the International Polar Year.
- **Who** – A partnership of the University of Alberta Libraries, University of Waterloo Canadian Cryospheric Information Network, OCUL Scholars Portal, Fisheries and Oceans Canada – Integrated Science Data Management, and NRC-CISTI.
- **Mandate** – To provide a sustainable research data management infrastructure, encompassing preservation and access, for polar (Arctic and Antarctic) science research and monitoring initiated from and taking place in Canada.
- **Intersects/interfaces** – Those government departments and agencies that have been designated by federal legislative documents (for example, the *Oceans Act*) to collect data for the purpose of understanding the environment and its living resources and ecosystems.
- **Priorities** – Data in scope but outside the Government data archival divisions will be a priority for the CPDN.

Canadian Astronomy Data Centre (CADC)
- **What** – One of the principal data archiving and data mining facilities worldwide for astronomical data.
- **Who** – A division of NRC.
- **Mandate** – Management, curation, preservation and access of all data for projects in which Canadian astronomers are involved (primarily academic).
- **Intersects** – The international astronomical community; the Canadian Space Agency.

Appendix 3: Bibliography

UK

Digital Curation Centre (DCC) UK. *A series of guides designed to build capacity and skills in research data management*. 2011 and 2012.
      How to Appraise and Select Research Data
      How to Cite Datasets and Link to Publications
      How to Develop a Data Management and Sharing Plan
      How to License Research Data
      How to Write a Lay Summary
http://www.dcc.ac.uk/resources/how-guides

Green, Ann, Stuart Macdonald and Robin Rice. 2009. Policy-making for Research Data in Repositories: A Guide. http://www.coar-repositories.org/files/guide.pdf

JISC. e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. 2003. http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

JISC. Data centres: their use, value and impact. September 2011. http://www.jisc.ac.uk/media/documents/publications/general/2011/datacentres.pdf

JISC. Digitisation in the UK: The Case for UK Framework. November 2005. www.jisc.ac.uk/publications/programmerelated/2005/pub_digi_uk.aspx

Hey, Tony ed., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009.

Hey, Tony and Anne E. Trefethen. *The UK e-Science Core Programme and the Grid*. http://users.ecs.soton.ac.uk/ajgh/FGCSPaper.pdf

Hey, Tony and Anne Trefethen. *The Data Deluge: An e-Science Perspective*. 2003. http://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf

Hey, Tony, and Jessie Hey. "e-Science and Its Implications for the Library Community." *Library Hi Tech* 24, no. 4 (2006): 515–28. http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2380240404.html

Maron , Nancy L, Jason Yun and Sarah Pickle. *Sustaining our digital future: Institutional Strategies for Digital Content*. January 29, 2013. JISC, Strategic Content Alliance. http://www.sr.ithaka.org/research-publications

Maron, Nancy L, and Matthew Loy. *Funding for Sustainability: How Funders' Practices Influence the Future of Digital Resources*. June 2011. www.sr.ithaka.org/research-publications/funding-sustainability-how-funders-practices-influence-future-digital

STFC. *Building an Open Data Infrastructure for Science: Turning Policy into Practice*. November 2012. www.ambafrance-uk.org/IMG/pptx/5_AngloFrenchBicarregui.pptx

STFC. *e-Science Department Strategy 2011*. http://www.stfc.ac.uk/e-science/resources/pdf/stfce-sciencedepartmentstrategy2011.pdf

STFC *e-Science Annual Review*. Daresbury Laboratory and Rutherford Appleton Laboratory. 2009-10. http://www.stfc.ac.uk/e-Science/resources/PDF/e-science_AnnRep_2010.pdf

Whyte, Angus and Jonathan Tedds. *Making the Case for Research Data Management*. DCC Briefing Paper. September 2011. http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm

The Royal Society. *Science as an open enterprise*. June 2012.
http://royalsociety.org/policy/projects/science-public-enterprise/report/

USA

ACLS. *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. 2006.
http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf

Edwards, Paul N., Steven J. Jackson, Geoffrey C. Bowker and Cory P. Knobel. Understanding Infrastructure: Dynamics Tension and Design. *Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures"*. February 2007.
http://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/UnderstandingInfrastructure2007.pdf;jsessionid=2EB84DB8B71BC66EAFFBD60D0CB5E2C8?sequence=3

Kahn, Scott D. "On the Future of Genomic Data". *Science* 331, 728 (2011); pp 728-729.
DOI: 10.1126/science.1197891

National Academies Press. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. 2009. http://www.nap.edu/catalog.php?record_id=12615

National Academies Press. *Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop*. 2010. http://www.nap.edu/catalog.php?record_id=12916

National Academies Press. *The Future of Scientific Knowledge Discovery in Open Networked Environments: Summary of a Workshop*. 2012.http://www.nap.edu/catalog.php?record_id=18258

National Institutes of Health. *Data and Informatics Working Group Draft Report to The Advisory Committee to the Director*. June 15, 2012.
http://acd.od.nih.gov/Data%20and%20Informatics%20Working%20Group%20Report.pdf

National Science Foundation. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure*. January 2003.
http://www.nsf.gov/od/oci/reports/atkins.pdf

National Science Foundation. *Cyberinfrastructure Framework for 21st Century Science and Engineering*. May 2012. http://www.nsf.gov/od/oci/cif21/CIF21Vision2012current.pdf

National Science Foundation. *Cyberinfrastructure for 21st Century Science and Engineering. Advanced Computing Infrastructure Vision and Strategic Plan*. February 2012.
http://www.nsf.gov/pubs/2012/nsf12051/nsf12051.pdf

National Science Foundation. *Cyberinfrastructure Framework for the 21st Century. A vision and strategy for software for science, engineering and education*. 2012.
http://www.nsf.gov/pubs/2012/nsf12113/nsf12113.pdf

NITRD. *Five-year strategic plan for the Federal Networking and Information Technology Research and Development (NITRD) Program*. July 2012. http://www.nitrd.gov/Publications/index.aspx

NITRD. *Harnessing the Power of Digital Data for Science and Society*. January 2009.
http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

OSTP Policy Memorandum. *Increasing Access to the Results of Federally Funded Scientific Research*.
February 22, 2013. http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

President's Council of Advisors on Science and Technology (PCAST). *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*. 2010. Report to the President and Congress. http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf

President's Council of Advisors on Science and Technology (PCAST). *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*. 2010. Report to the President and Congress. January 2013.
http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd2013.pdf

Weller, Martin. *The Digital Scholar: How Technology Is Transforming Scholarly Practice*. 2011.
http://www.bloomsburyacademic.com/view/DigitalScholar_9781849666275/book-ba-9781849666275.xml

Australia

Australian Code for the Responsible Conduct of Research. 2007.
http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf

Australian Government. *An Australian e-Research Strategy and Implementation Framework*. Final Report of the e-Research Coordinating Committee. April 2006.
http://ncris.innovation.gov.au/Documents/eRCCReport.pdf

Australian Government. *Platforms for Collaboration. Summary Investment Plan*.
http://ncris.innovation.gov.au/Documents/PfCInvPlansum.pdf

Australian Government. *Platforms for Collaboration Final Investment Plan*.
http://ncris.innovation.gov.au/Documents/PfCInvPlanFinal.pdf

Australian National Data Service (ANDS). Diverse guides and policies.
http://ands.org.au/guides/index.html

O'Brien, Linda. *The Changing Scholarly Information Landscape: Reinventing Information Services to Increase Research Impact*. June 2010. Proceedings ELPUB2010 – Conference on Electronic Publishing.
http://elpub.scix.net/cgi-bin/works/Show?111_elpub2010

Wolski, Malcolm, Joanna Richardson, Mark Fallu, Robyn Rebollo, Joanne Morris. *Developing the Discovery Layer in the University Research e-Infrastructure*. 2011.

Wolski, Malcolm, Joanna Richardson and Robyn Rebollo. "Building an Institutional Discovery Layer for Virtual Research Collections". *D-Lib Magazine* May/June 2011, Volume 17, Number 5/6.
http://www.dlib.org/dlib/may11/wolski/05wolski.html

Europe

ASPIRE: A Study on the Prospects of the Internet for Research and Education. Middleware and Managing Data and Knowledge in a Data-rich World. Funded by FP7. September 2012. http://www.terena.org/activities/aspire/docs/ASPIRE-data.pdf

CERN. *Open Federated Identity Management for Research Collaborations.* April 2012. CEWRN 2012-006. https://cdsweb.cern.ch/record/1442597/files/CERN-OPEN-2012-006.pdf

Dallmeier-Tiessen S, Darby R, Gitmans K, Lambert S, Suhonen J, Wilson M (2012).*Opportunities for Data Exchange. Compilation of results on drivers and barriers and new opportunities.* 09 July 2012 for FP7. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf

Deutsche Forschungsgemeinschaft (DFG). *Taking Digital Transformation to the Next Level: The Contribution of the DFG to an Innovative Information Infrastructure for Research.* July 2012. http://www.dfg.de/download/pdf/foerderung/programme/lis/strategy_paper_digital_transformation.pdf

DFG Committee on Scientific Library Services and Information Systems; Subcommittee on Information Management. *Recommendations for Secure Storage and Availability of Digital Primary Research Data.* Status 26 June 2008. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901_en.pdf

European Commission. *ICT Infrastructures for eScience.* Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. 2009 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0108:FIN:EN:PDF

European Commission. *Open Infrastructures for Open Science. Horizon 2020 Consultation Report.* http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/open-infrastructure-for-open-science.pdf

European Commission. *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. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

European Commission. *Advancing Technologies and Federating Communities: A Study on Authentication and Authorisation Platforms For Scientific Resources in Europe.* 2012. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf

European Commission. *Development of impact measures for e-Infrastructures.* 2012. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/impact-study-final-report.pdf

European Commission. *Developing World-class Research Infrastructures for the European Research Area (ERA). Report of the ERA Expert Group.* http://ec.europa.eu/research/infrastructures/pdf/ri_era-expert-group-0308_en.pdf

European Commission. *G8+O5 Global Research Sub-group on Data.* Draft Report, 28 October 2011. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/g8.pdf

European Commission. *Knowledge without Borders. GÉANT 2020 as the European Communications Commons.* October 2011. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/geg-report.pdf

European Commission. *Global Virtual Research Communities*.
ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/e-infrastructure/dge01224j-brochurea5-global-virtual-research-com-low_en.pdf

European Commission. *Financing a Software Infrastructure for Highly Parallelised Codes* — IDC Final Report for the DG Information Society of the European Commission. July 2011.
http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/fisofi4hpc.pdf

European Commission. *Digital Object Identifiers and Uniqur Author Identifiers to Enable Services for Data Quality Assessment, Provenance and Access*. Draft Final Report. October 2011.
http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/digoiduna.pdf

European Commission. E-Infrastructure Reflection Group. *e-IRG Roadmap*. 2012. http://www.e-irg.eu/images/stories/publ/e-irg_roadmap_2012-final.pdf

European Commission. E-Infrastructure Reflection Group. *e-IRG Cloud Computing for research and science: a holistic overview, policy, and recommendations*. Final Version. 30 October 2012.
http://www.e-irg.eu/images/stories/dissemination/e-irg_cloud_computing_paper_v.final.pdf

European Commission. Franco Accordino. *Digital Science and its impact on Scientific Societies*. 2010.
www.cs.uu.nl/groups/AD/panel-**DigitalScience**.pdf

European Commission. Celina Ramjoué. *Towards a European Policy on Open Access*. November 2012.
http://www.unica-network.eu/sites/default/files/2012-11_27_UNICA-Celina.pdf

European Commission. EGI. *Linking digital resources across Eastern Europe for European science and Innovation*. 2012. http://www.terena.org/activities/development-support/epe2012/presentations/11.pdf

European Commission. *Strategic Plan for e-infrastructure. (European Grid Infrastructure EGI)*. EGI-InSPIRE. 2010. https://documents.egi.eu/public/RetrieveFile?docid=1098&version=3&filename=EGI-1098-D230-final.pdf

European Commission. *Report on Integration of Data and Publications*. October 17, 2011. Susan Reilly, Wouter Schallier, Sabine Schrimpf, Eefke Smit, Max Wilkinson.
http://www.libereurope.eu/sites/default/files/ODE-ReportOnIntegrationOfDataAndPublication.pdf

EU and UK Department for Culture, Media and Sport. *Dynamic Action Plan for the EU co-ordination of digitisation of cultural and scientific content*. 2005.
http://www.minervaeurope.org/publications/dap/dap.pdf

GRDI2020. *Final Roadmap Report, Global Research Data Infrastructures: The Big Data Challenges*. Submitted to the EU, February 2012. http://www.grdi2020.eu/Repository/FileScaricati/e2b03611-e58f-4242-946a-5b21f17d2947.pdf

Siena. *Roadmap on Distributed Computing Infrastructure for e-Science*. May 2012.
http://www.sienainitiative.eu/

Other International

McKinsey & Company. *Big Data: The next frontier for innovation, competition, and productivity*, James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. May 2011.
http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

OECD. *Principles and Guidelines for Access to Research Data from Public Funding*. OECD, 2007. Available at: www.oecd.org/dataoecd/9/61/38500813.pdf

Canada

Backgrounder 2011. *Canadian Research Data Summit 2011: Canadian Research Data Summit. Mapping the Data Landscape*. http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/docs/Summit_Backgrounder.pdf

Canadian Polar Data Network (CPDN) Governance Charter. 2013. http://polardatanetwork.ca/wp-content/uploads/CPDN_Governance.pdf

CARL. The Canadian National Collaborative Data Infrastructure Project. Final Report. January 2012. Prepared by Lynn Copeland, CNCDI Project Coordinator, and Martha Whitehead, Chair, CARL Data Management Subcommittee. http://carl-abrc.ca/uploads/pdfs/carl_cncdi_final_report.pdf

Canada's Digital Environment for Research, Innovation and Education. A Submission under the Digital Economy Strategy Consultation by Canadian Digital Media Network, Canadian Research Knowledge Network, Canadian Council of CIOs, CANARIE Inc. and Compute Canada. June 28, 2010. http://www.canarie.ca/templates/about/publications/docs/DES_Submission_E.pdf

Canadian Association for Public Data Use (CAPDU). *Consultation on the Future Role of the National Archives of Canada and the National Library of Canada*. December 1998. http://datalib.library.ualberta.ca/data/capdu-english-consultation-submissions.pdf

Comprehensive Brief on Open Access to Publications and Research Data in Canada. http://www.science.gc.ca/default.asp?lang=en&n=2360F10C-1

Digital Infrastructure Discussions: A précis of the discussions leading up to the Digital Infrastructure Summit, June 13-14, 2012 at the University of Saskatchewan, Saskatoon. 2012.

English, John. *The Role of the National Archives of Canada and the National Library of Canada*. A Report submitted to the Honourable Sheila Copps. 1999. http://datalib.library.ualberta.ca/data/englishreport1999.pdf

Genome Canada. Meeting Report from the Bioinformatics & Computational Biology Workshop. Toronto, Ontario, Canada. December 5-6, 2011. http://www.genomecanada.ca/medias/pdf/en/bioinformatics-meeting-report.pdf

Humphrey, Chuck. *Preserving Research Data in Canada*. Blog, December 2012. http://preservingresearchdataincanada.net/

Library and Archives Canada. *Canadian Digital Information Strategy* (Consultation Version). October 2007. http://datalib.library.ualberta.ca/data/CDISfinalreport.pdf

Library and Archives Canada. *Canadian Digital Information Strategy (CDIS): Final Report of consultations with stakeholder communities 2005 to 2008*. Published February 2010 http://datalib.library.ualberta.ca/data/CDIS_FinalReport_eng_REVISED_Final.pdf

Mapping the Data Landscape: Report of the 2011 Canadian Research Data Summit. December 2011. Produced by the Research Data Strategy Working Group. http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/docs/data_summit-sommet_donnees/Data_Summit_Report.pdf

Research Data Canada website. http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/index.html

Royal Society of Canada. *Data Policy and Barriers to Data Access in Canada: Issues for Global Change Research*. 1996. A Discussion Paper by the Data and Information Systems Panel of the Canadian Global Change Program. Available from the Royal Society of Canada.

Shearer, Kathleen and Diego Argáez. 2010. *Addressing the Research Data Gap: A Review of Novel Services for Libraries*. Report to CARL. http://carl-abrc.ca/uploads/pdfs/library_roles-final.pdf

Mapping the Data Landscape: Report of the 2011 Canadian Research Data Summit. http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/events/data_summit_2011/index.html

Pearce, Nick, Martin Weller, Eileen Scanlon, Sam Kinsley. "Digital Scholarship Considered: How New Technologies Could Transform Academic Work". *Technology & Social Media in Education* (Special Issue, Part 2), 2010, 16(1). http://www.ineducation.ca/article/digital-scholarship-considered-how-new-technologies-could-transform-academic-work

Social Sciences and Humanities Research Council. *National Research Data Archive Consultation Phase One: Needs Assessment Report*. May 2001. http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/da_phase1_e.pdf

Social Sciences and Humanities Research Council. Final Report. National Data Archive Consultation. *Building Infrastructure for Access to and Preservation of Research Data*. Submitted by the NDAC Working Group to the Social Sciences and Humanities Research Council of Canada and the National Archivist of Canada. June 2002. http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/da_finalreport_e.pdf

Stewardship of Research Data in Canada: A Gap Analysis. 2009. http://publications.gc.ca/collections/collection_2009/cnrc-nrc/NR16-123-2008E.pdf

Strong, David F. and Peter B. Leach. *National Consultation on Access to Scientific Research Data* (NCASRD). Final Report, January 31, 2005. http://datalib.library.ualberta.ca/data/NCASRDReport_e.pdf

Appendix 4: Definitions

What	Definition	Source
Data	Qualitative or quantitative statements or numbers that are (or assumed to be) factual. Data may be raw or primary data (e.g., direct from measurement), or derivative of primary data, but are not yet the product of analysis or interpretation other than calculation.	Science as an Open Enterprise
Big data	"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. This definition assumes that as technology advances over time, the size of datasets that qualify as big data will increase. Also the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).	McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity
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.	Mapping the Data Landscape 2011 Summit
Metadata	Metadata—"data about data"—contains information about a dataset. This may state why and how it was generated, who created it and when. It may also be technical, describing its structure, licensing terms and standards it conforms to.	Science as an Open Enterprise
Open data	Open data is data that meets the criteria of intelligent openness. Data must be accessible, useable, assessable and intelligible.	Science as an Open Enterprise
Semantic data	Data that are tagged with particular metadata that can be used to derive relationships between data.	Science as an Open Enterprise
Curation	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.	JISC e-Science Curation Report
Archiving	A curation activity that ensures that data is properly selected stored and can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.	JISC e-Science Curation Report

Preservation	An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.	JISC e-Science Curation Report
Data management	A process involving a broad range of activities for handling data, from administrative to technical aspects.	Mapping the Data Landscape 2011 Summit
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.	Mapping the Data Landscape 2011 Summit
Research Data Management Infrastructure (RDMI)	RDMI is the configuration of staff, services and tools assembled to support data management across the research lifecycle and more specifically to provide comprehensive coverage of the stages making up the data lifecycle. It can be organized locally and/or globally to support research data activities across the research lifecycle.	Blog. Humphrey, Chuck
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.	Mapping the Data Landscape 2011 Summit
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.	Mapping the Data Landscape 2011 Summit
Data interoperability	The structuring of data in such a way that diverse datasets can be integrated.	Mapping the Data Landscape 2011 Summit
Dataset	Any organized collection of data.	Mapping the Data Landscape 2011 Summit
Metadata (1)	Metadata is descriptive or contextual information which refers to or is associated with another object or resource. This usually takes the form of a structured set of elements that describe the information resource and assist in the identification, location and retrieval of it by users, while facilitating content and access management.	Data Curation Centre Website, UK
Metadata (2)	Metadata is data about data… More generally, information consists of semantic tags applied to data. Metadata consists of semantically tagged data that are used to describe data. Metadata can be organized in a schema and implemented as attributes in a database.	Hey and Trefethen: The Data Deluge

Descriptive metadata	• Enables identification, location and retrieval of information resources by users, often including the use of controlled vocabularies for classification and indexing and links to related resources.	Data Curation Centre Website, UK
Technical metadata	Describes the technical processes used to produce, or required to use a digital object.	Data Curation Centre Website, UK
Administrative metadata	• Used to manage administrative aspects of the digital object such as intellectual property rights and acquisition. Administrative metadata also documents information concerning the creation, alteration and version control of the metadata itself. This is sometimes known as meta-metadata.	Data Curation Centre Website, UK
Use metadata	Manages user access, user tracking and multi-versioning information.	Data Curation Centre Website, UK
Preservation metadata	Documents actions that have been undertaken to preserve a digital resource such as migrations and checksum calculations.	Data Curation Centre Website, UK
Grid	Any distributed infrastructure that is federated to combine resources from multiple organizations managed by different administrative domains. The Grid aims to coordinate the sharing of resources in a dynamic and multi-institutional setting to provide additional functionality beyond its constituent parts: brokering, workflow coordination, integration of computing and storage. In order for this to happen, interoperability and standards need to be defined at various levels: for resource access, for coordination and business logic, for data storage and management, for network access and so forth.	European Commission, Advancing Technologies and Federating Communities
e-Science (1)	The term e-Science is used to represent the increasingly global collaborations—of people and of shared resources—that will be needed to solve the new problems of science and engineering.	Hey and Trefethen: The Data Deluge
e-Science (2)	Computationally intensive science that is carried out in highly distributed network environments, or science that uses immense datasets that require grid computing; the term sometimes includes technologies that enable distributed collaboration, such as the Access Grid.	Wikipedia
e-Science (3)	E-Science is not a new scientific discipline in its own right: e-Science is shorthand for the set of tools and technologies required to support collaborative, networked science. The entire e-Science infrastructure is intended to empower scientists to do their research in faster, better and different ways.	Hey, Tony, and Jessie Hey. "e-Science and Its Implications for the Library Community." *Library Hi Tech* 24, no. 4 (2006): 515-28.

Digital scholarship	Incorporates: • building a digital collection of information for further study and analysis; • creating appropriate tools for collection-building; • creating appropriate tools for the analysis and study of collections; • using digital collections and analytical tools to generate new intellectual products; and • creating authoring tools for these new intellectual products, either in traditional forms or in digital form.	Our Cultural Commonwealth
Cyber-infrastructure	• Grids of computational centres • Comprehensive libraries of digital objects • Well-curated collections of scientific data • Online instruments and vast sensor arrays • Convenient software toolkits	In Our Cultural Commonwealth referring to the Aitkins Report (UK 2003)
e-Research infrastructure	Comprises the ICT assets, facilities and services that support research within institutions and across national innovation systems, and that enable researchers to undertake excellent research and deliver innovation outcomes.	Rys Francis Presentation
Cyber-infrastructure	Those layers that sit between base technology (a computer science concern) and discipline-specific science. The focus is on value-added systems and services that can be widely shared across scientific domains, both supporting and enabling large increases in multi- and interdisciplinary science while reducing duplication of effort and resources—e.g., including hardware, software, personnel, services and organizations.	Aitkins Report
Cloud computing	A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms and services are delivered on demand to external customers over the Internet. Key elements: • it is a specialized distributed computing paradigm; • it is massively scalable; • it can be encapsulated as an abstract entity that delivers different levels of services to customers outside the Cloud; • it is driven by economies of scale; and • the services can be dynamically configured (via virtualization or other approaches) and delivered on demand. Effectively it provides access to powerful computing capabilities without investing in new infrastructure, training new personnel or licensing new software.	GRDI 2020

Scientific data infrastructure	What is required to enable researchers to create, store and share the data resulting from their experiments, and to find, access and process the data they need	European Commission, Advancing Technologies and Federating Communities
	raw data collected, produced during experiments, surveys or observations of different phenomena (according to an initial research model); the data is consequently analyzed and the findings published. A preservation process is needed during all these stages; **structured data** and datasets resulting from data filtering and processing (supporting some particular formal model); **published data** organized in a way to support a scientific theory and/or research results; and **data publishing** to support research consolidation, integration, and openness.	
Workflows	A workflow is a precise description of a scientific procedure—multi-step processes to coordinate multiple tasks, acting like a sophisticated script. Each task represents the execution of a computational process. Data output from one task is consumed by subsequent tasks according to a predefined graph topology that "orchestrates" the flow of data.	Hey: *The Fourth Paradigm*. Goble and de Roure. p. 138.
Data file formats	Preferred formats are formats designated by a data repository for which it guarantees that they can be converted into data formats that will remain readable and usable. Usually, these are the de facto standards employed by that particular community.	Quoted in Policy-making for Research Data in Repositories: A Guide 2009