Preserving Research Data: where are we now?

ACCOLEDS 2007

Chuck Humphrey
University of Alberta

Outline

- Today's drivers to preserve research data
- Threads traceable to earlier studies and how they have shaped today's dialogue on data archiving
  - **To establish a national data management strategy**
  - **To determine how big the problem of research data loss really is**
  - **To distribute data archiving functions**
  - **To provide access to publicly funded data**
- The "open" movements and how they are shaping data environments
- Partnerships in the evolving distributed repository model

Drivers to preserve research data

- New concerns raised about the stewardship of publicly funded resources, including research data.
  - **For example, new policies in health (NIH and CIHR)**
- A confluence of “open” movements bringing together software, research outputs and data.
- The emergence of e-Science and its emphasis on access to scientific data.
  - **Short-term access does not require preservation but long-term access will not be possible without preservation.**
- The transformation of the Digital Library movement from digitizing content to accessing content.

Threads from earlier studies

A National Data Management Strategy

- ***Data Policy and Barriers to Data Access in Canada: Issues for Global Change Research*** by the Data and Information Systems Panel of the Canadian Global Change Program, 1996

  "Current Canadian data collection and management strategies can be characterized as localized and confused… No national data management strategy exists to ensure a rational data collection scheme, standards of data management, equitable data access, nor the long-term preservation of those data that are collected." (p. 14)

Threads from earlier studies

A National Data Management Strategy

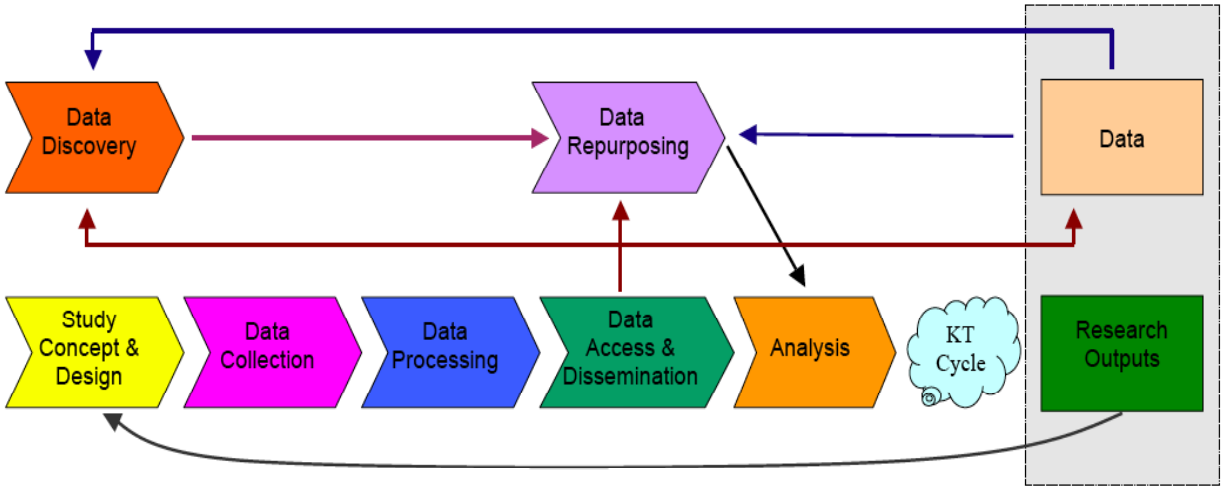
- In 1998, the John English consultation on the future role of the National Archives and the National Library.
  - **CAPDU submission recognized in the final report**

> **"RECOMMENDATION: WE ENDORSE THE CANADIAN ASSOCIATION OF PUBLIC DATA USERS PROPOSAL FOR A NATIONAL DATA MANAGEMENT STRATEGY IN WHICH THE NATIONAL ARCHIVES AND THE NATIONAL LIBRARY PLAY A FACILITATIVE ROLE. THE TWO INSTITUTIONS SHOULD PLAY A PARTNERSHIP ROLE WITH SUCH A DATA ARCHIVE AND COORDINATE THE FEDERAL GOVERNMENT'S RELATIONSHIP WITH SUCH AN ARCHIVE." (p. 18-19)**

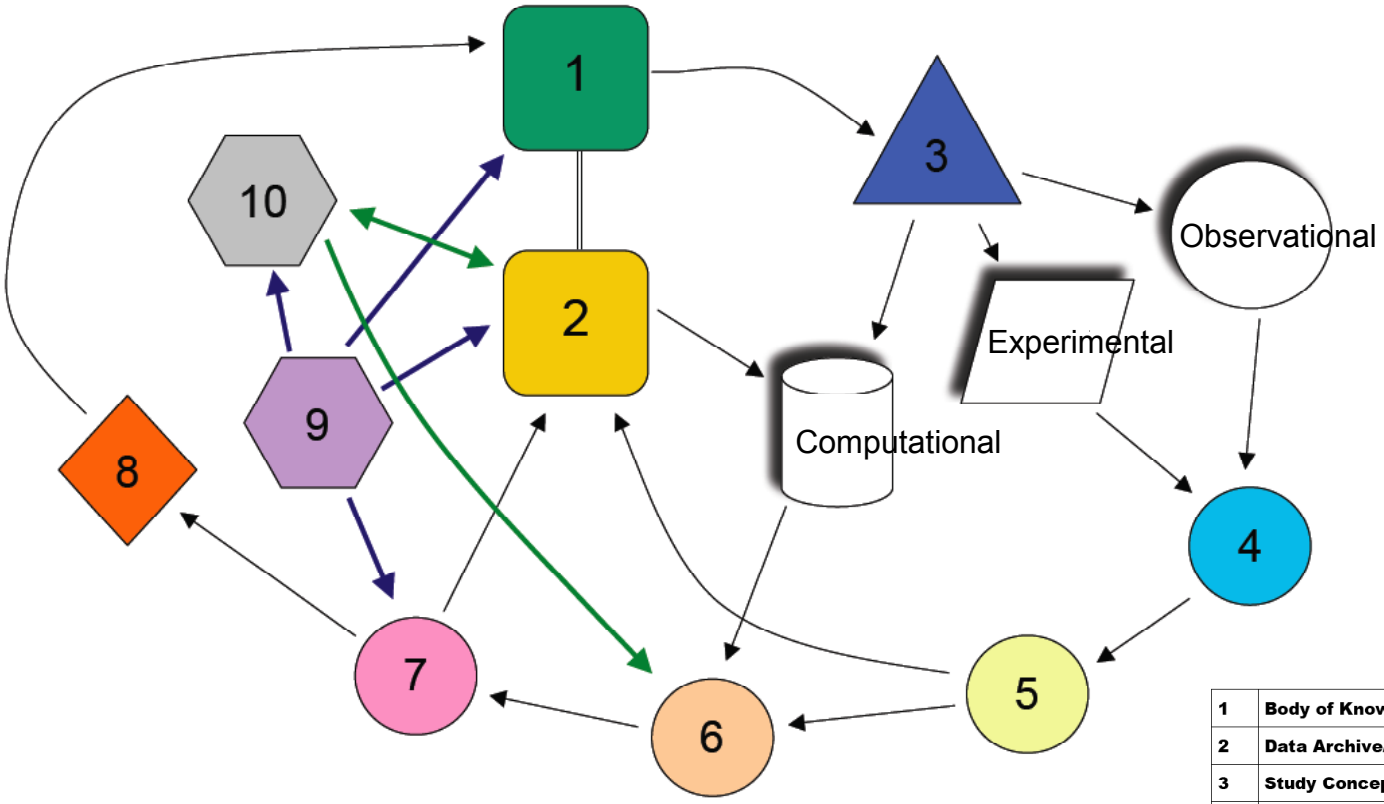
Research Life Cycle Model

- The thread calling for a data management strategy has evolved into the research life cycle model for thinking about the stages needed to represent a comprehensive management strategy.

One example of a research life cycle model. Source: **To Stand the Test of Time** http://www.arl.org/bm~doc/digdatarpt.pdf



Representation of the Research Life Cycle



1	Body of Knowledge	
2	Data Archive/Repository	
3	Study Concept / Design	
4	Data Collection or Capture	
5	Data Processing / Dissemination	
6	Computational Models or Analysis	
7	Data Visualization / Analysis Files	
8	Research Outputs / Knowledge Transfer	
9	Data Discovery	
10	Data Repurposing	

Source: Humphrey, *Metadata-driven Solutions*, Open Data Foundation Meeting, September 13-15, 2007 at http://opendatafoundation.org/

Where we are today

A National Data Management Strategy

- While Canada is still without a national data management strategy, attention to this issue has surfaced in the following events.
  - **Janet Halliwell's introductory remarks at the National Consultation on Access to Scientific Research Data (NCASRD) quoted from the report of the Data and Information Systems Panel of the Canadian Global Change Program and challenged the assembly to put this shortcoming in Canada to rest. Sections in the NCASRD final report mention both data management plans as a best practice and international leadership in data management.**
  - **The Canadian Digital Information Strategy contains an action item (1.3.2) that also addresses data management plans as part of Canadian funding programs.**

Where we are today

A National Data Management Strategy

- The recent focus on data management strategies has been at the individual research project level around data production, dissemination and preservation. The national infrastructure to support individual research plans has not been clearly identified yet, although discussions have suggested that the solution should be at the university or discipline level.
- Work continues in articulating the life cycle model as it applies to managing and preserving research data.

Threads from earlier studies

Determining the severity level of lost research data

- The SSHRC and the National Archives joint consultation (NDAC) consisted of two phases. Phase 1, which lasted from October 2000 to June 2001, was asked to demonstrate a need for data archiving in Canada. Prior to Phase 1's report, only anecdotal evidence existed about the risk level facing research data in Canada. Determining actual numbers of files at risk is a difficult task, as noted in **Data Policy and Barriers to Data Access in Canada.**

> **"Understandably, the published literature is remarkably silent on the subject of lost data collections, and as no comprehensive inventory of datasets exists in Canada, the identification of lost data is a matter of inside knowledge and luck." (p. 15)**

The data risk level in Canada

Three studies were conducted in conjunction with Phase 1 of the NDAC that provided evidence about the level to which Canadian research data are at risk.

Source: Humphrey, *Preserving Research Data: A Time for Action*, **Preservation of Electronic Records: New Knowledge and Decision-making**, Ottawa: The Canadian Conservation Institute (2004), pp. 83-90.

The data risk level in Canada

Study 1: a gap-analysis of existing mandates and practices of national institutions.

Findings:

- **The vast majority of academic and non-academic research data fall outside the current [ca. 2001] interpretation and execution of the mandates of the National Library and National Archives.**
- **No other Canadian institutions have a national mandate or the resources to address the current level of need for preserving research data.**

The data risk level in Canada

Study 2: a follow-up study to an investigation first conducted twenty years ago by the now defunct Machine-Readable Archives.

Findings:

- **Data from 3 out of 110 studies could be found without contacting the original principal investigators directly for further details.**
- **The 3 studies for which data were found all were deposited in the United States with the Inter-university Consortium for Political and Social Research (ICPSR).**

Conclusion:

- **The risk of data loss is very high without an institution with the specific mandate to preserve research data.**

The data risk level in Canada

Study 3: a survey of researchers receiving a standard research grant from the SSHRC between 1998 and 2000.

Findings:

- **Without a recognized institution responsible for preserving research data, researchers do not know where or how to archive the data from their research, even if they would like to see the data preserved.**
- **For the vast majority of researchers in this study, archiving data is an unknown activity in conducting research.**
- **The survey of researchers who received a standard research grant from the SSHRC provides evidence that around 550 out of every 1,000 projects results in the creation or use of data files and/or databases.**

Where we are today

Determining the severity level of lost research data

- No national accounting system has been developed to establish an on-going record of the number of data files with research value that are at risk. The facts gathered in the 2000-01 NDAC remain our best indicators on data loss.
- A recent study by Carol Perry does corroborate evidence from the 2000-01 NDAC findings that preserving data is not part of the research culture in the social sciences and humanities.

University Affairs
Affaires universitaires

June-July 2007

Data storage policy can't be enforced

by Moira MacDonald

Survey shows 70 percent of researchers are unaware they're required to archive data collected from SSHRC-funded research

Social Sciences and Humanities Research Council grant recipients have been required for the last 17 years to permanently archive their research data, but a recent study found most recipients it surveyed were unaware of the requirement.

The study, by University of Western Ontario graduate student Carol Perry, found that more than 70 percent of respondents were unaware of SSHRC's "Research Data Archiving Policy." Some of those who did know about the policy were concerned that it contradicted their own university's requirement to destroy data after a fixed period of time.

SSHRC's policy requires that "all research data collected with the use of SSHRC funds must be preserved and made available for use by others within a reasonable period of time."

The policy, in existence since 1990, lists 11 sites across the country where researchers can archive their data if their own university cannot.

Chad Gaffield, SSHRC's president, told University Affairs that while all recipients should be aware of SSHRC's archiving policy, it was not yet enforceable. "It is impossible to try to police this in an aggressive way," Dr. Gaffield said. "We know the structures are not in place."

Through her studies and work in information services at the University of Guelph, "I pretty much knew that people weren't archiving their data," said Ms. Perry, a candidate for a master's degree in information and library sciences. "In Canada there [isn't] a well-defined mechanism in place [to archive research data], and there's no Canada-wide policy, although most institutions are discussing it and recognizing the need to go toward setting up a data repository."

Carol Perry at U of Guelph says Canada doesn't have a well-defined mechanism to archive research data.
Photo: Krys Mooney

Threads from earlier studies

A distributed model of data archiving functions

- Phase 1 of the National Data Archive Consultation made reference to a “research data archiving function” rather than an institution.

> **“For the purpose of this report, a research data archiving function is defined as preserving, managing and making publicly accessible digital information structured through methodology for the purpose of producing new knowledge.” (p. 4)**

NDAC: Phase One Needs Assessment Report
http://www.sshrc.ca/web/about/publications/da_phase1_e.pdf

Threads from earlier studies

A distributed model of data archiving functions

- Phase 2 of the NDAC, which was given the green light to proceed in June 2001, examined existing international models for providing data archiving services and recommended three possible models for Canada.
- A tension exists in the report between an institutional solution and a network of partnerships. Pages 10-13 provide a potential list of partners in a "National Research Data Archive network." They include: university data services, Canadian archival institutions, international connections, CANARIE, management frameworks (e.g., REB's), partner institutions (e.g., LAC), preservation services for other agencies.

Threads from earlier studies

A distributed model of data archiving functions

- The three models recommended in the final report:
  - **Through federal legislation, create a National Research Data Archive Network as a modified version of a Separate Statutory Agency.**
  - **Create a National Research Data Archive Network under the auspices of the SSHRC.**
  - **Create a Special Agency within LAC to preserve research data.**

Where we are today

A distributed model of data archiving functions

- Recent discussion tends to support a distributed model that takes advantage of institutional repositories on local campuses. While universities have been identified as only one group among the partnerships needed for a national data archiving service, no dialogue has been initiated to bring all possible partners to the table. The proposal in the NCASRD report was to create an entity, called Data Force, to create such a forum.

Threads from earlier studies

Data as a Public Investment

- OECD Ministerial Declaration on Access to Research Data from Public Funding
  - **dated January 30, 2004**
  - **Canada is one of 33 co-signers**
- Premise of the declaration: publicly funded research data should be openly available to the maximum extent possible.
- Subsequent selective implementations in the U.S. and the U.K.
- In Canada, this contributed to forming the National Consultation on Access to Scientific Research Data.

Threads from earlier studies

Data as a Public Investment

- NCASRD was an NRC, NSERC, CFI and CIHR sponsored consultation in 2004-05 "*to help Canada maximize the value received from its publicly funded natural and medical sciences research by recommending an appropriate framework and guidelines, which will facilitate open and long-term access to data coming from that research.*" http://ncasrd-cnadrs.scitech.gc.ca/home_e.shtml
- In the NCASRD final report, a proposal is made to establish a dedicated national infrastructure, known as Data Canada, to oversee the implementation of the report's recommendations. Data Canada would support other institutions involved in archiving data by providing national leadership. http://ncasrd-cnadrs.scitech.gc.ca/NCASRDReport_e.pdf

Where we are today

Data as a Public Investment

- CIHR introduced a new policy regarding open access. The policy is fairly specific about research outputs, but is vague about preserving data and only addresses access to publication-related data.

> **"CIHR has decided to limit this policy to peer-reviewed journal publications and publication-related biomedical research data, which is typically deposited into public databases as a condition of publication. CIHR … will explore broadening the policy to include research materials and other research data in the future. "**
>
> **http://www.cihr-irsc.gc.ca/e/34846.html**

Threads from more recent studies

It takes a research community to preserve its data

- An ARL-NSF Consultation was held in the fall of 2006 to provide direction to NSF's Cyber-infrastructure program, which is known as e-Science in Europe.
- The final report, "*To Stand the Test of Time: Long-term stewardship of digital data sets in science and engineering*," argues for new partnerships among domain scientists, librarians, and data scientists to manage digital data collections better.
- A new NSF program arising from this report is NSF 07-601, "Sustainable Digital Data Preservation and Access Network Partners."

Threads from more recent studies

It takes a research community to preserve its data

- The Canadian Digital Information Strategy was released for public review in October of this year. This document contains a set of principles to guide digital information policies in the cultural, government and scientific sectors. The preservation of scientific data is included in this report. (see: http://www.collectionscanada.gc.ca/cdis/index-e.html)

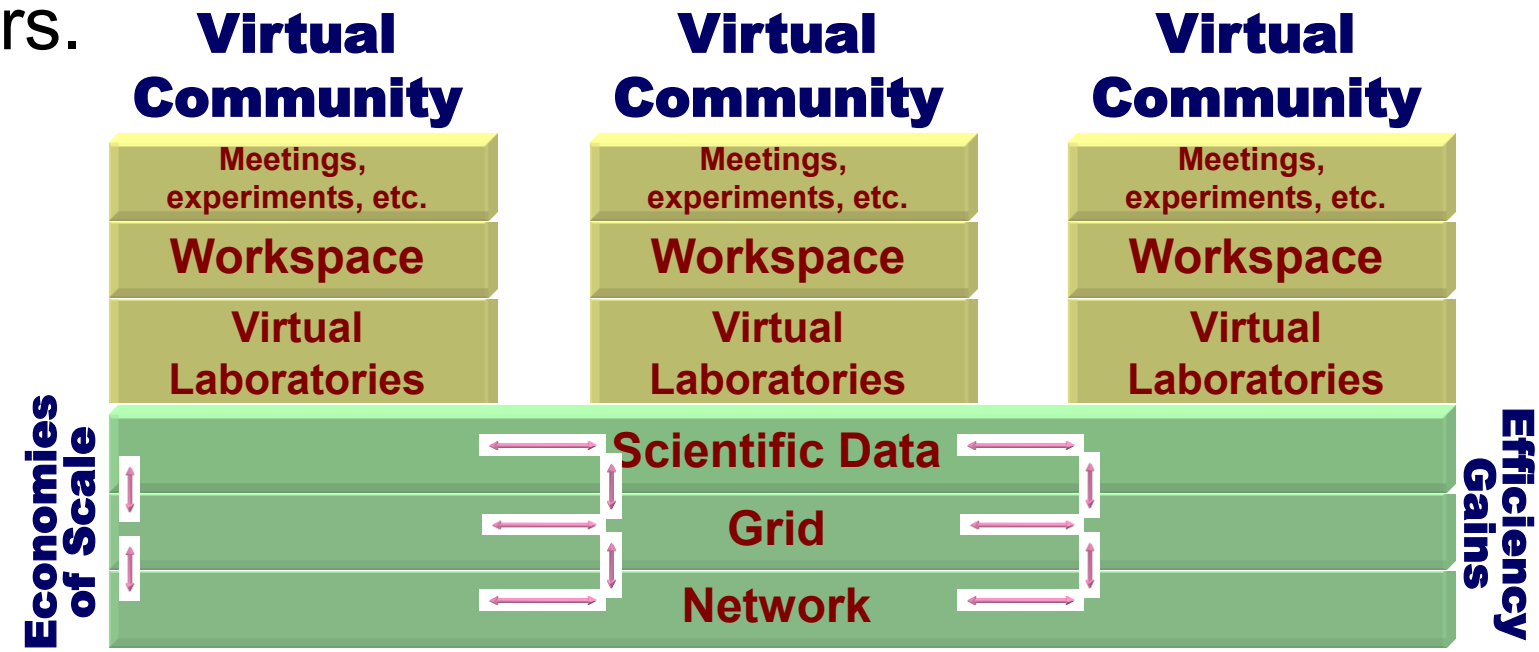
Where we are today

It takes a research community to preserve its data

- We are creating preservation-quality metadata in Canada. The Research Data Centre Network received a CFI award to develop metadata and supporting tools for the confidential data it houses.
  - **This project is based on a life-cycle model of research and examines the capture and preservation of metadata throughout the stages of research.**
- Statistics Canada has two internal DDI projects: one in DLI and another with a group of author divisions.
- OCUL's Data group (DINO) received funding from OntarioBuys to produce DDI metadata for data holdings in Ontario universities. This project is known as ODESI.

Open movements shaping research environments

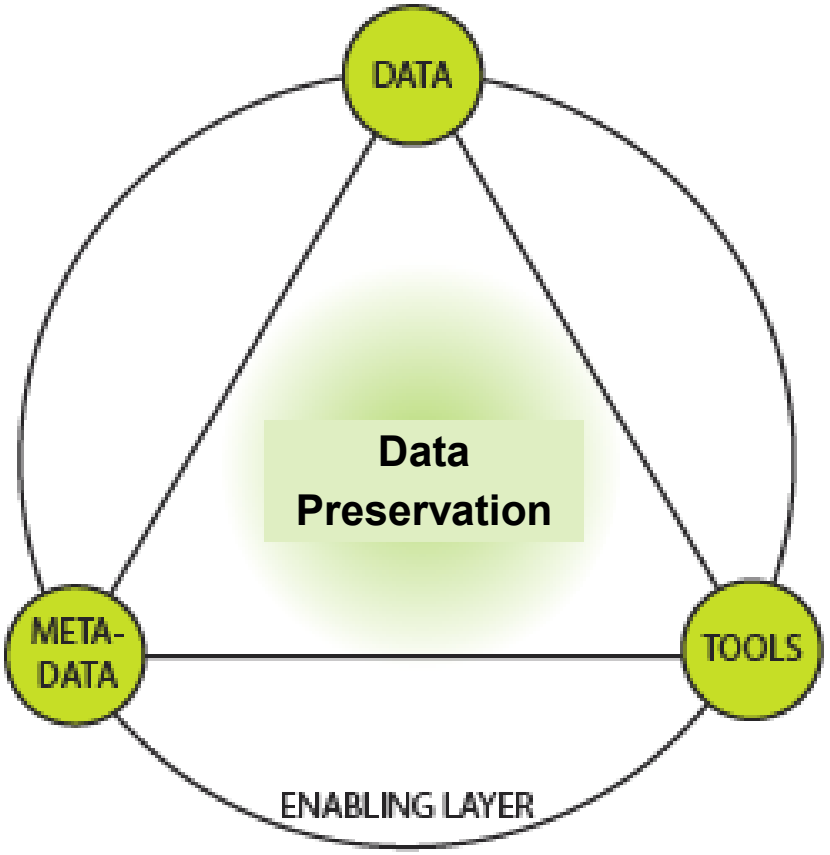
- Investments in e-Science or cyber-infrastructure underlie the environments being developed for researchers. This infrastructure consists of layers.



Source:Ulf Dahlsten, "Building a global virtual research community," http://internationaldataforum.org

An infrastructure framework

- Another framework is represented in the proposal for the European Research Observatory for the Humanities and Social Sciences (EROHS).

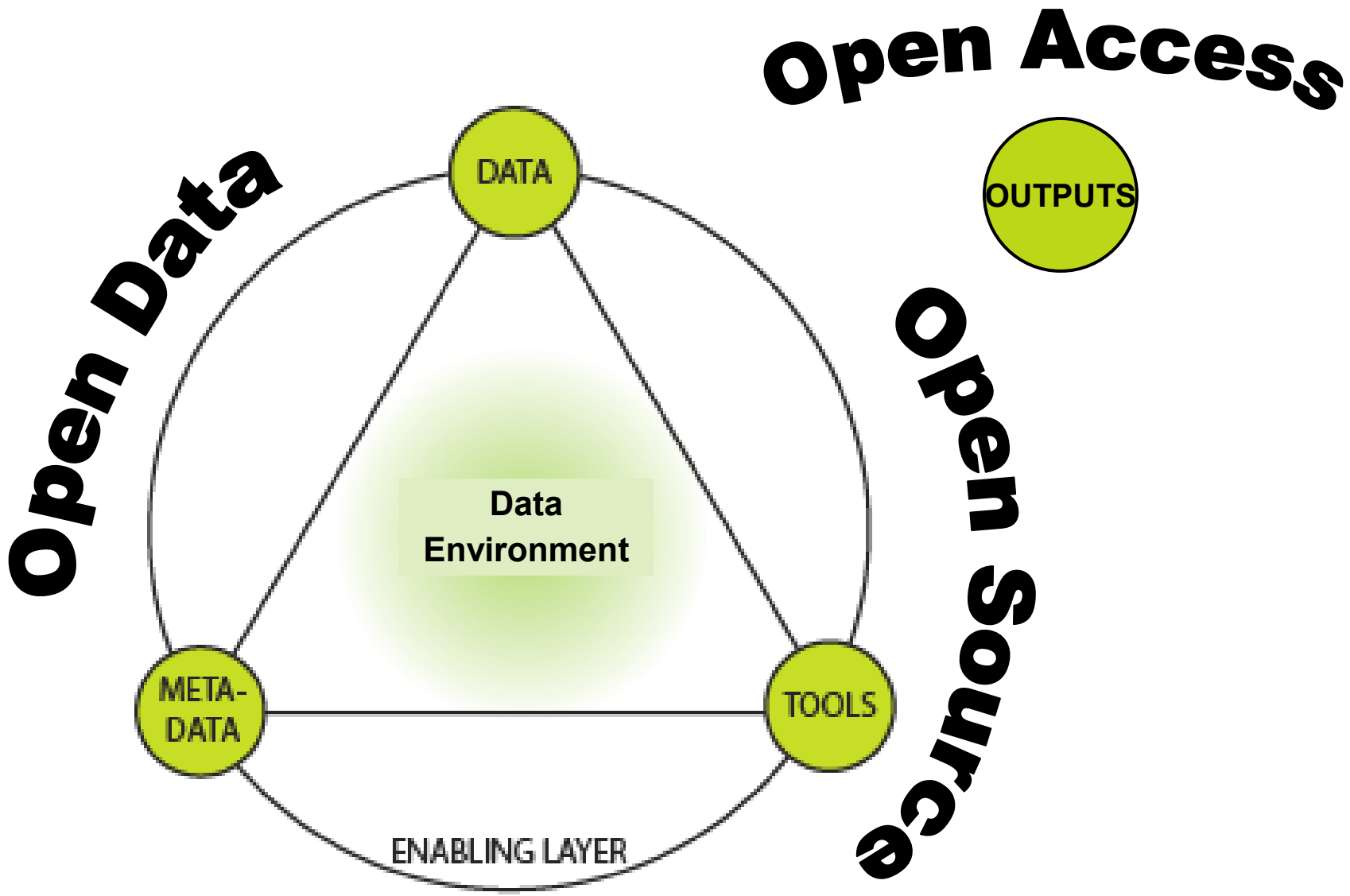


Bjørn Henrichsen

Chair, ESFRI Social Science and Humanities Roadmap Working Group

IASSIST May 24 2006

Open movements



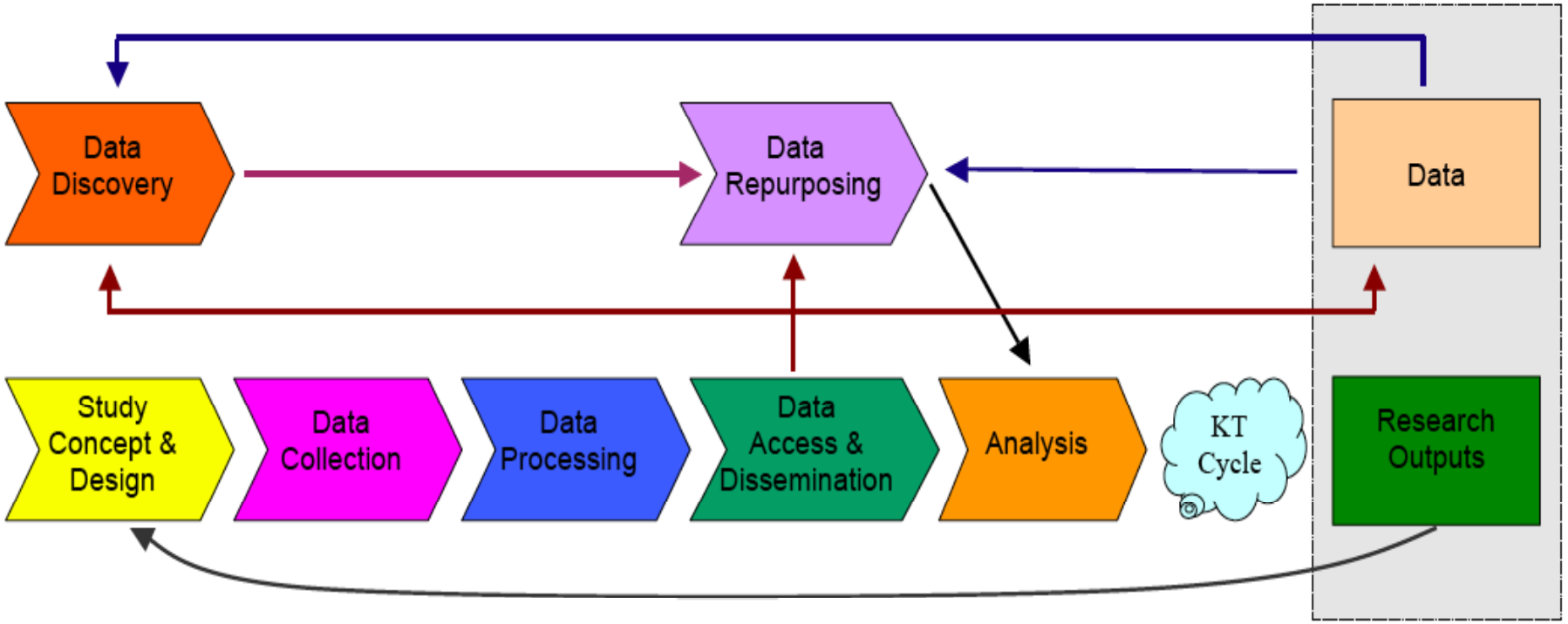
Implementation strategies

- There will be a mix of strategies to implement large, complex cyber-infrastructure frameworks, such as the one articulated by the ESFRI Social Sciences and Humanities Roadmap Working Group.
- Among the mix of strategies will be national and international initiatives from research councils and agencies funding infrastructure development.
- Open data environments also have a role to play in these implementation strategies.

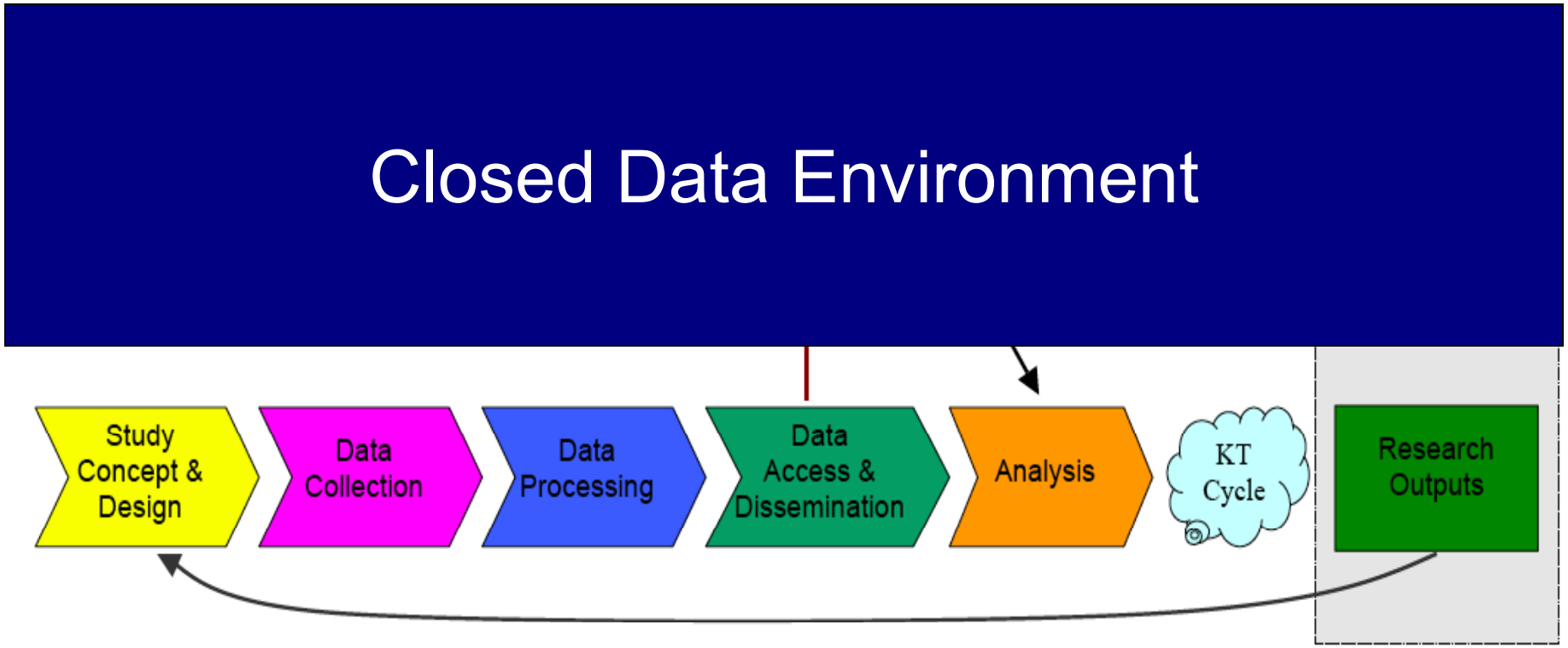
Data Environments

	Are the metadata open?	
Are the data open?	No	Yes
No	Closed Environment	Discovery Environment
Yes	File Retrieval Environment	Repurposing Environment

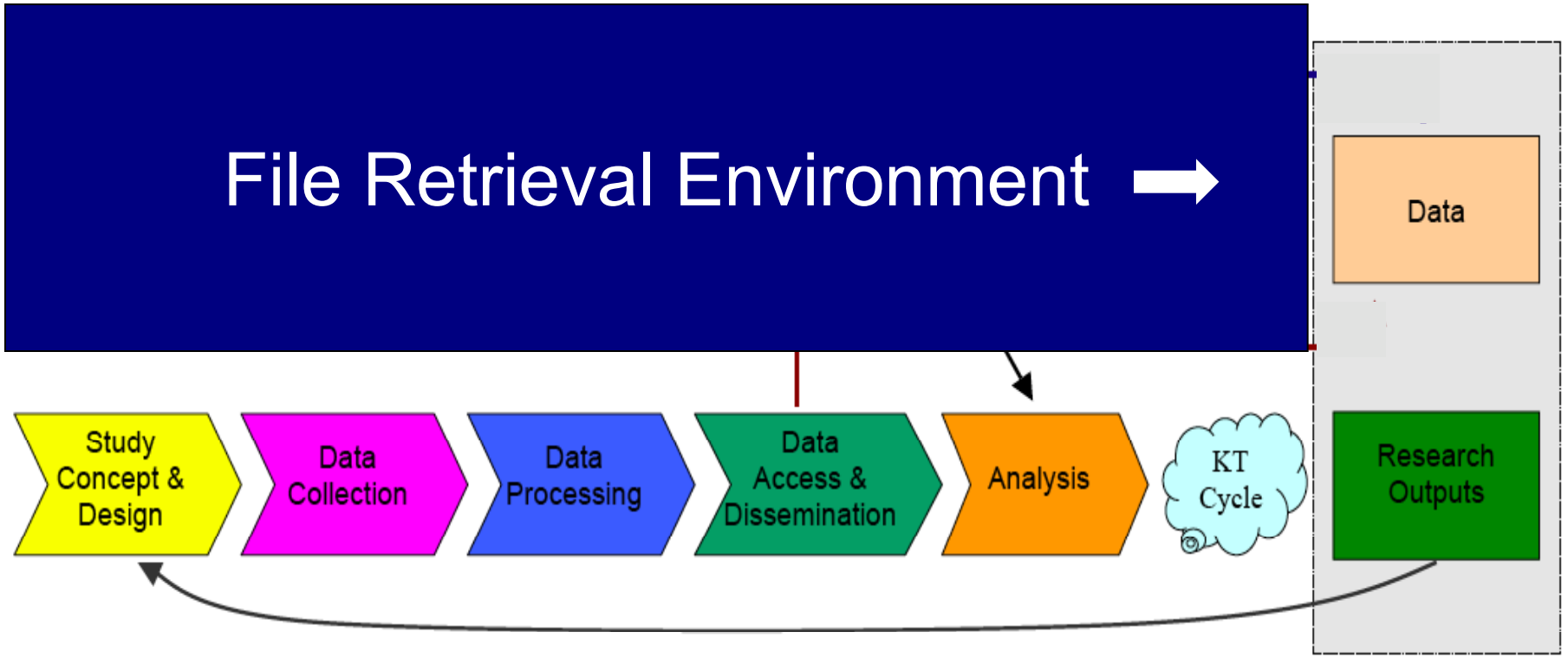
Life Cycle Model of Research Knowledge Creation



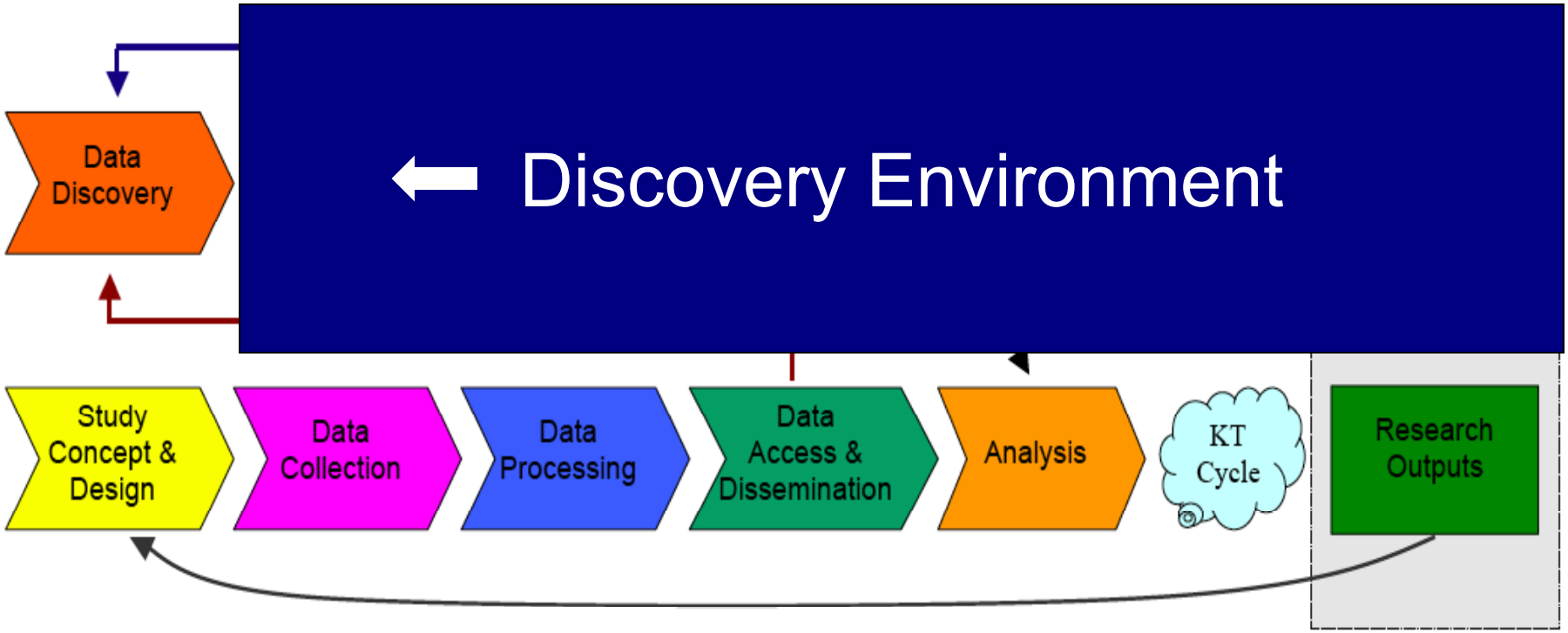
No Open Metadata : No Open Data



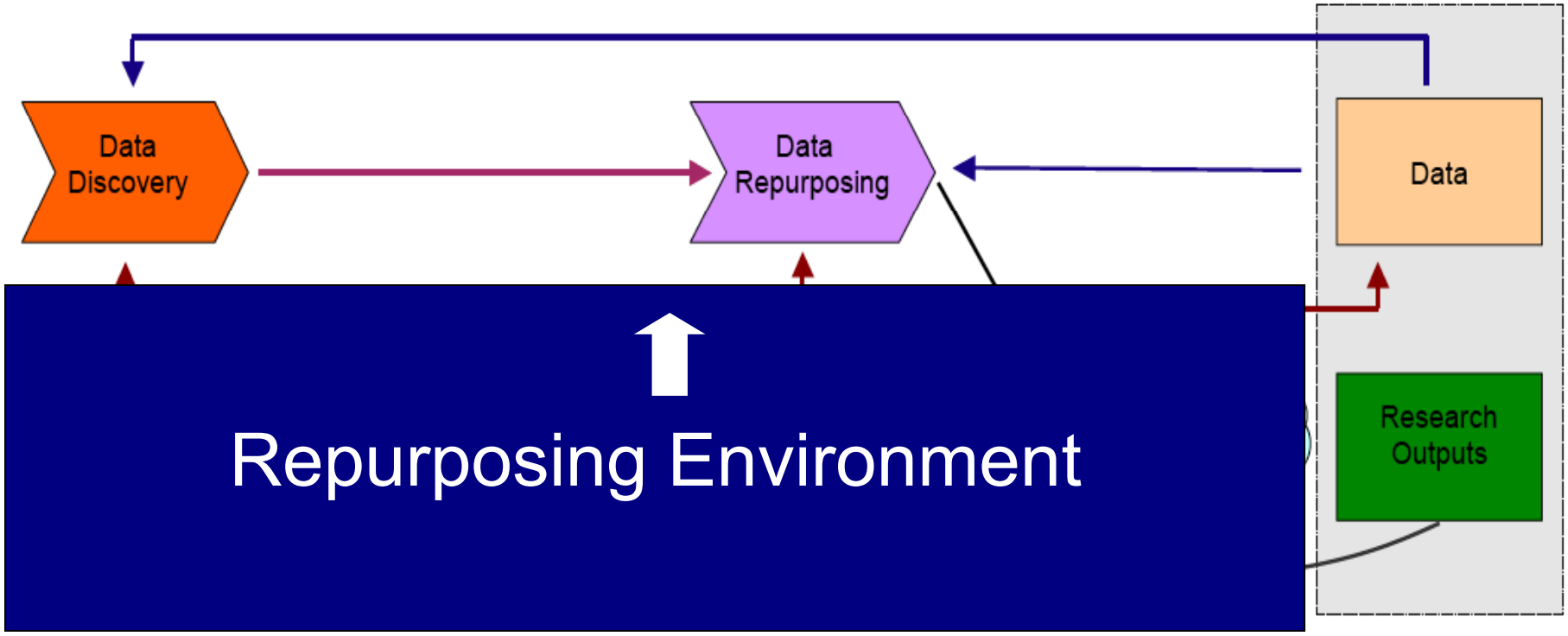
No Open Metadata : Open Data



Open Metadata : No Open Data

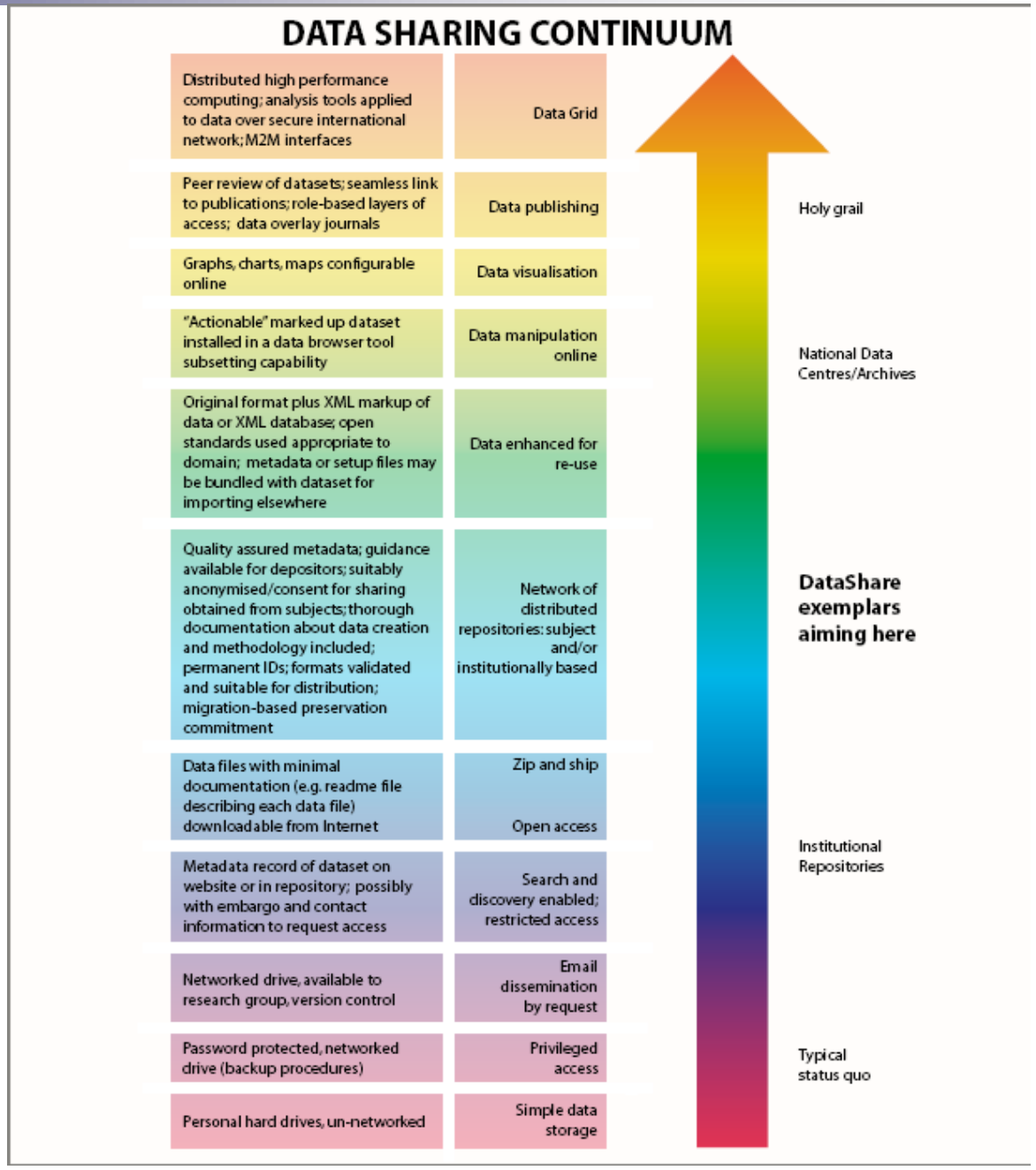


Open Metadata : Open Data



These four data environments can also be mapped onto Robin Rice's data sharing continuum. The closed data environment is at the bottom of the scale; the file retrieval environment is in the light blue section; the discovery environment is the mid-blue section without data access and repurposing is the top half of the chart.

http://www.disc-uk.org/docs/ECDLposter.pdf



The “open” digital library agenda

- The agenda of the international digital library movement in the 1990’s was to create digital content.
- Ten years later, the agenda has shifted to capturing digital content on campuses and to providing access through institutional repositories.
  - **This movement was also driven by the open access and open publishing movements to offset the cost of expensive scientific journal subscriptions.**
  - **Research data were a casual part of the initial repository agenda and tended to be rolled in with other “digital objects,” but not given serious attention.**

Where we are today

Open movements and cyber-infrastructure

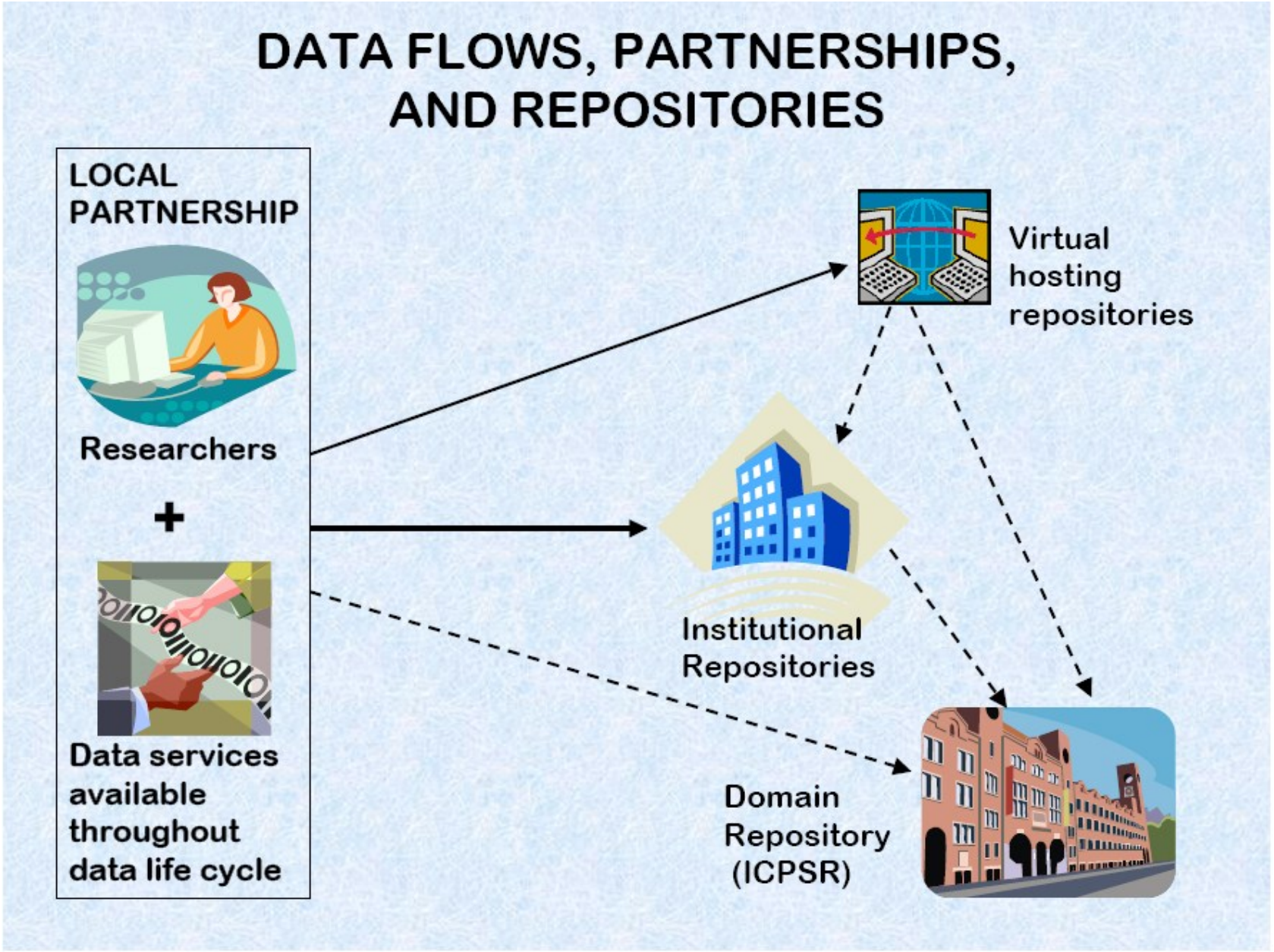
- Funding councils are beginning to invest in cyber-infrastructure for the social sciences.
  - **CESSDA just received 2.7M euros for research infrastructure development under the European Commission Seventh Framework Programme (FP7).**
  - **In the U.S., Data-PASS is a data preservation alliance funded by the Library of Congress National Digital Information Infrastructure Preservation Program (NDIIPP). Data-PASS partners consist of the ICPSR, Odum Institute, Roper Center, Henry A. Murray Research Archive, Harvard-MIT Data Center and NARA.**
  - **In Canada, CFI funded the Research Data Centre project to create DDI 3.0 metadata and tools; and OntarioBuys has invested in the OCUL ODESI DDI project.**

Where are we headed?

- In Canada, we lack a collective direction and a coordination of efforts. As Kevin Schurer (UKDA) said at the 2007 ICPSR OR meeting, “They [Canada] can’t get their act together.”
- I wrote a letter to Chad Gaffield (SSHRC) on August 30, 2007, which I subsequently sent to the Chair of the SSHRC Standing Committee on Research Support, asking him to initiate a process to build a national social sciences data strategy that stands on the principles articulated in the CDIS report. The intent would be to bring together stakeholders who would develop a larger plan for social science infrastructure and would identify roles for the various stakeholders. This could be modeled after the UK Data Forum and the International Data Forum.

Where are we headed?

- Developments in establishing institutional repositories are happening at the institutional and regional levels across Canada. The ACCOLEDS submission to the recent COPPUL strategic plan proposed a regional project around open journal publishing, tables in articles and linkages to data sources.
- Ann Green proposed a model of weaving these various developments into a safety net (not her words but mine) for research data, which involves local data services on campuses.



Source: Ann Green, *Connecting the Dots among digital repositories, data services and social science researchers*, ICPSR OR Meeting, October 2007

Where are we headed?

- This approach is largely pragmatic and builds on related but not necessarily connected infrastructure and services emerging at the local institution level.
- Canada will still need a national data archiving service to facilitate research data preservation but the distributed nature of data preservation may be very dependent on local support.
- The results in Canada may be similar to the model envisioned in Europe (see the next slide), where institutions instead of nations form the hubs in the network.

Future European SSH infrastructure

Bjørn Henrichsen
IASSIST - Ann Arbor
May 24 2006