Data-Driven Research, Data Infrastructure Services, and Challenges for Information Science

Jian Qin

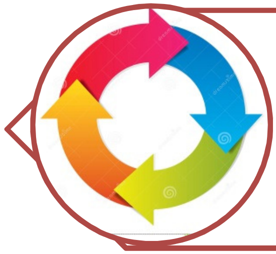
School of Information Studies

Syracuse University

Syracuse, New York, USA

XV ENANCIB, October 27, 2014, Belo Horizonte, Brazil

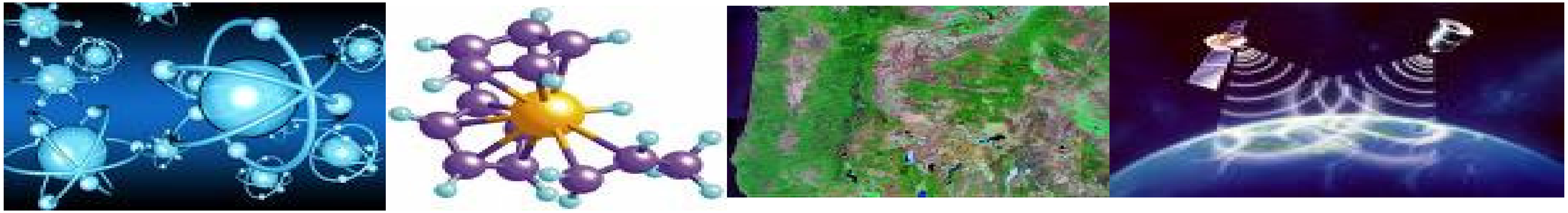
Three themes in this presentation



Data-driven research

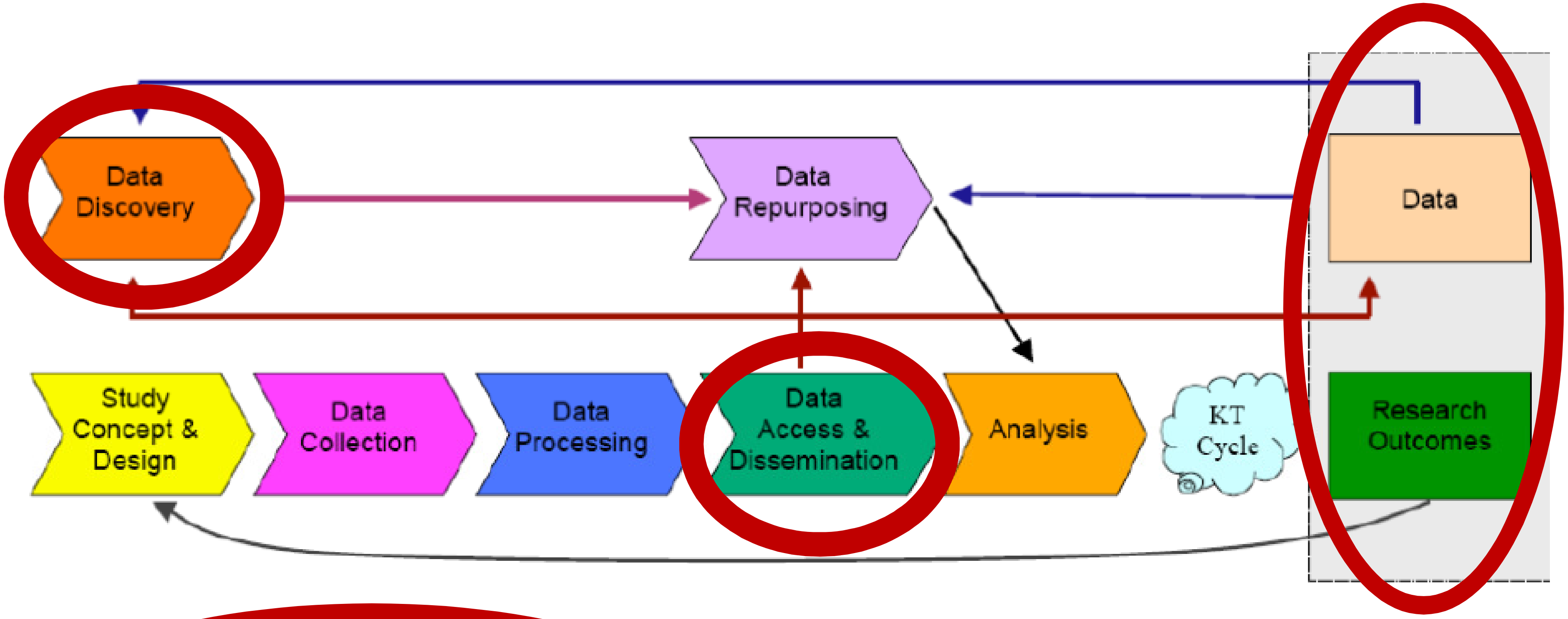
New territories of library and information services

Building capacities for data services

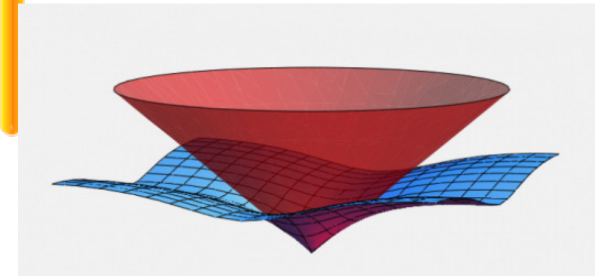


Data-driven research

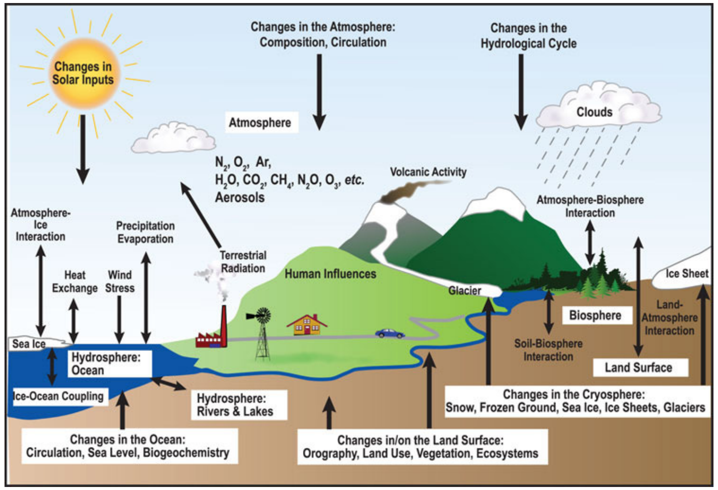
E-Science and the life cycle of research

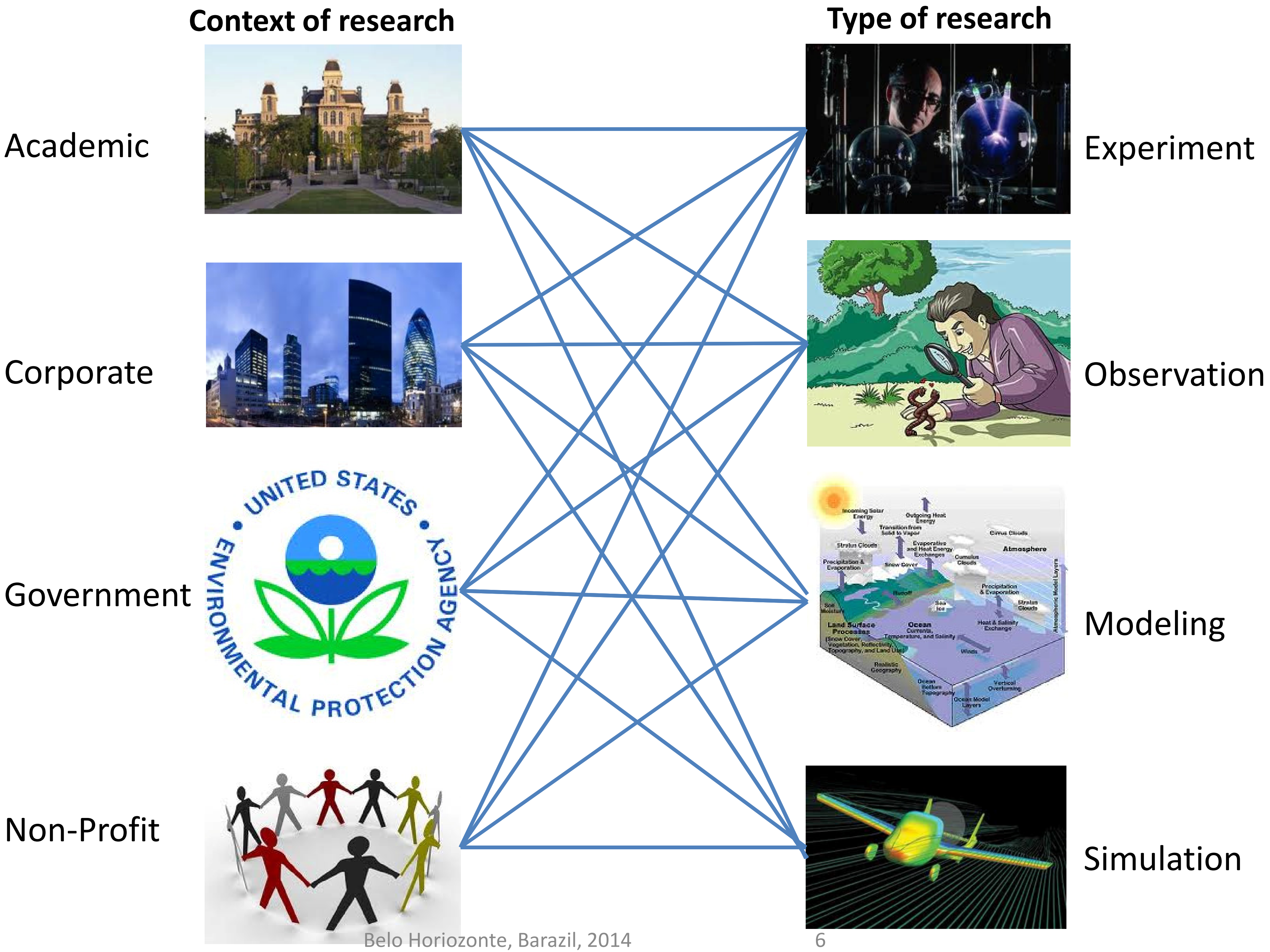


Where library and information professionals can contribute and make an impact



But what research?


Changes in the Atmosphere: Composition, Circulation
Changes in the Hydrological Cycle
Changes in Solar Inputs
Clouds
Atmosphere
N_2, O_2, Ar, H_2O, CO_2, CH_4, N_2O, O_3, etc.
Aerosols
Volcanic Activity
Atmosphere-Biosphere Interaction
Atmosphere-Ice Interaction
Precipitation Evaporation
Terrestrial Radiation
Human Influences
Heat Exchange
Wind Stress
Glacier
Ice Sheet
Biosphere
Land-Atmosphere Interaction
Sea Ice
Hydrosphere: Ocean
Soil-Biosphere Interaction
Land Surface
Ice-Ocean Coupling
Hydrosphere: Rivers & Lakes
Changes in the Cryosphere: Snow, Frozen Ground, Sea Ice, Ice Sheets, Glaciers
Changes in the Ocean: Circulation, Sea Level, Biogeochemistry
Changes in/on the Land Surface: Orography, Land Use, Vegetation, Ecosystems


Context of research

Type of research

Academic

Corporate

Government

Non-Profit

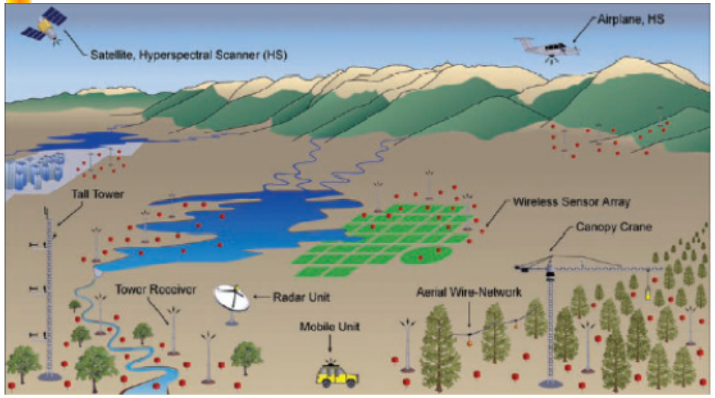
Experiment

Observation

Modeling

Simulation

A scenario of data collection



NSF. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Instruments that collect data

- Sensors
- Microwave towers
- Remote sensing

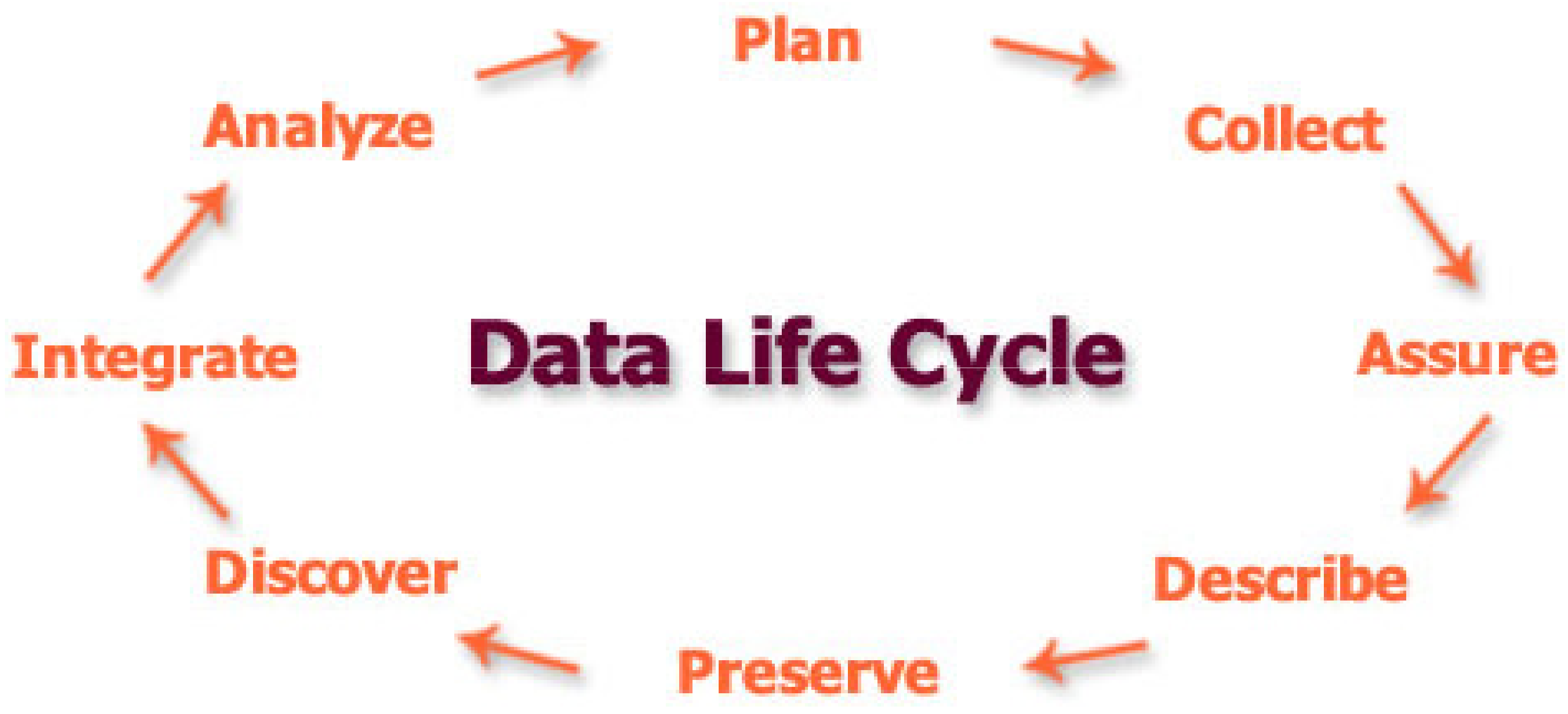
Level one processing:

- Formatting
- Calibrating
- Documenting
- Archiving (of raw data)
- Delivering (copies to research team)

Level two processing:

- Organizing data into appropriate data files and segments
- Converting metrics / measurements
- Delivering level two processing copies back to data archive

Research data life cycle



http://www.dataone.org/best-practices

Research data collections

Size

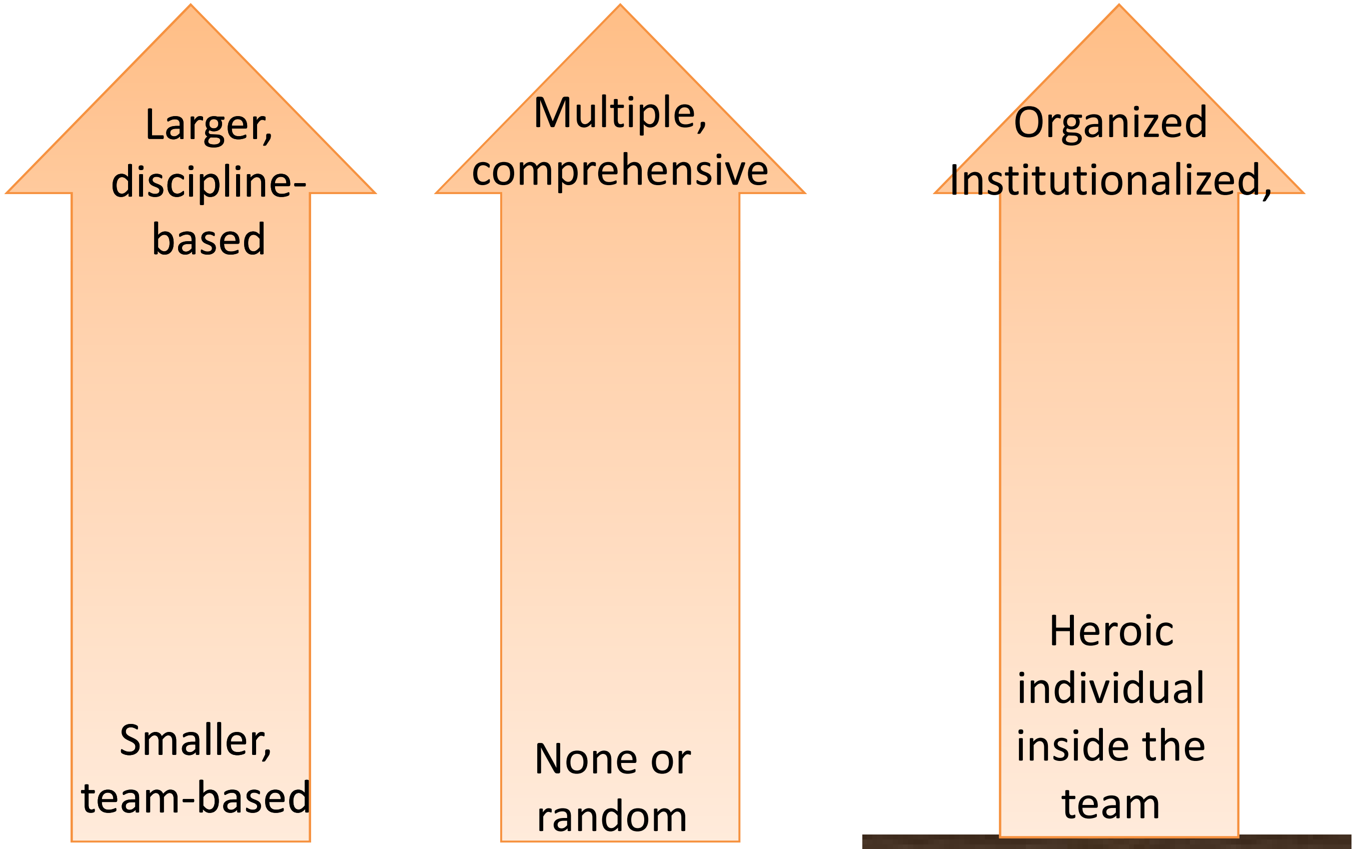
Metadata Standards

Management

Reference collection

Resource collection

Research collection



Larger, discipline-based

Smaller, team-based

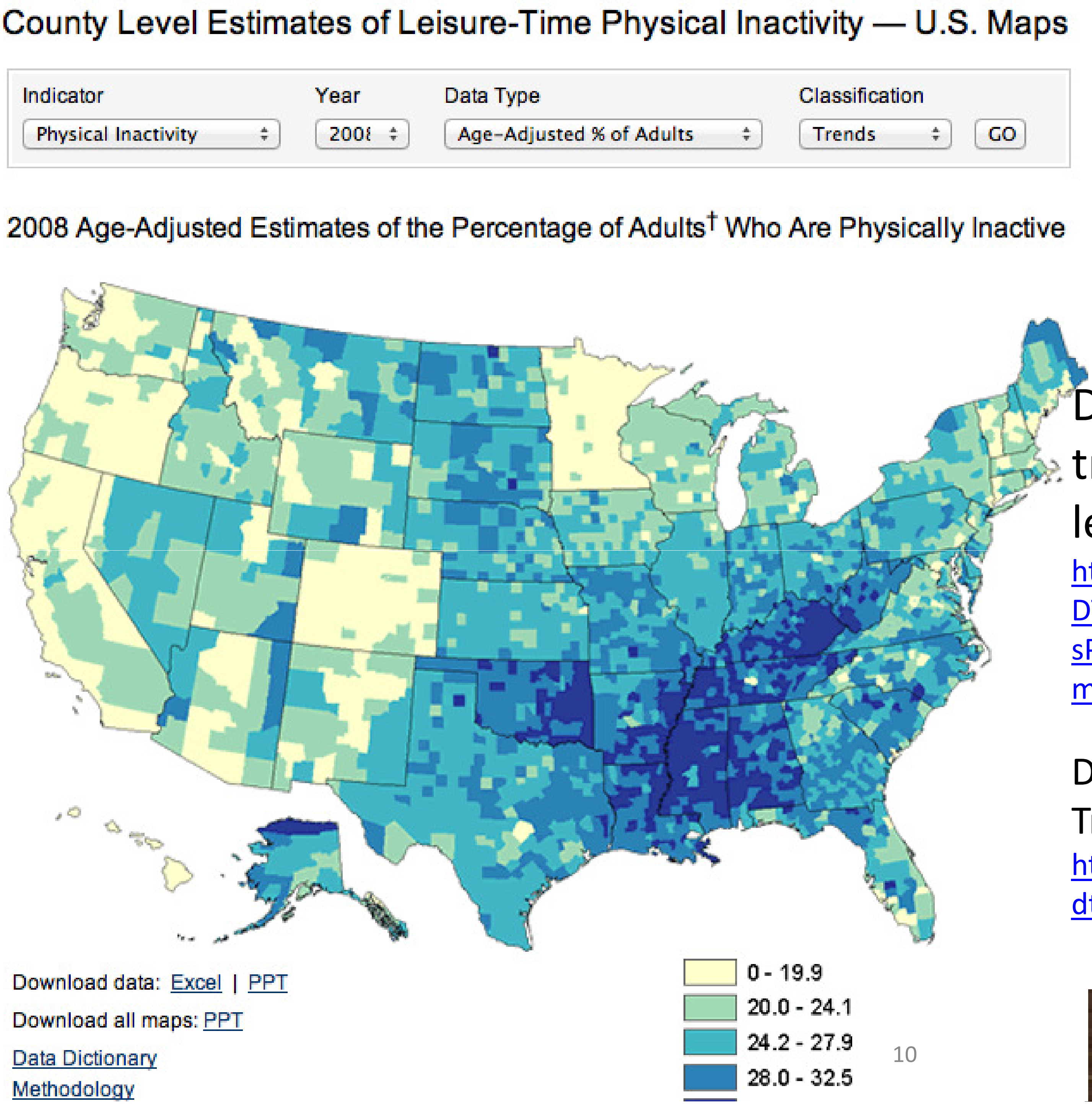
Multiple, comprehensive

None or random

Organized Institutionalized,

Heroic individual inside the team

Interactive data products



County Level Estimates of Leisure-Time Physical Inactivity — U.S. Maps

Indicator: Physical Inactivity

Year: 2008

Data Type: Age-Adjusted % of Adults

Classification: Trends

2008 Age-Adjusted Estimates of the Percentage of Adults[†] Who Are Physically Inactive

Download data: Excel | PPT

Download all maps: PPT

Data Dictionary

Methodology

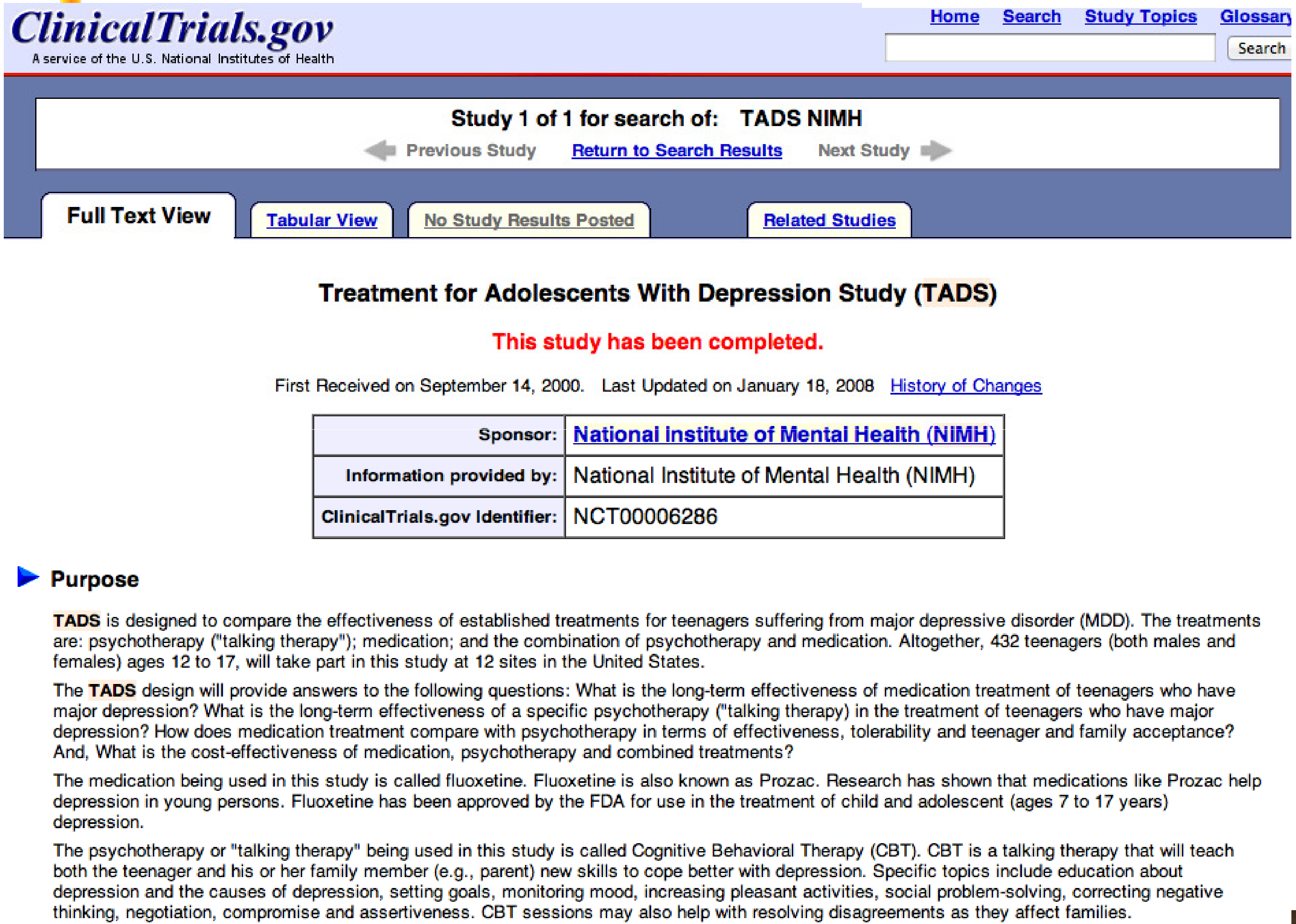
▭	0 - 19.9
🟩	20.0 - 24.1
▨	24.2 - 27.9
🟦	28.0 - 32.5

Diabetes data and trends—Country level estimates: http://apps.nccd.cdc.gov/DDT_STRS2/NationalDiabetesPrevalenceEstimates.aspx?mode=PHY ;

Diabetes Data & Trends home page: http://apps.nccd.cdc.gov/ddtstrs/default.aspx

Data registry

Clinical trials data management: http://www.clinicaltrials.gov/ct2/show/NCT00006286?term=TADS+NIMH&rank=1



Study 1 of 1 for search of: TADS NIMH

Treatment for Adolescents With Depression Study (TADS)

This study has been completed.

First Received on September 14, 2000. Last Updated on January 18, 2008 History of Changes

Sponsor:	National Institute of Mental Health (NIMH)
Information provided by:	National Institute of Mental Health (NIMH)
ClinicalTrials.gov Identifier:	NCT00006286

Purpose

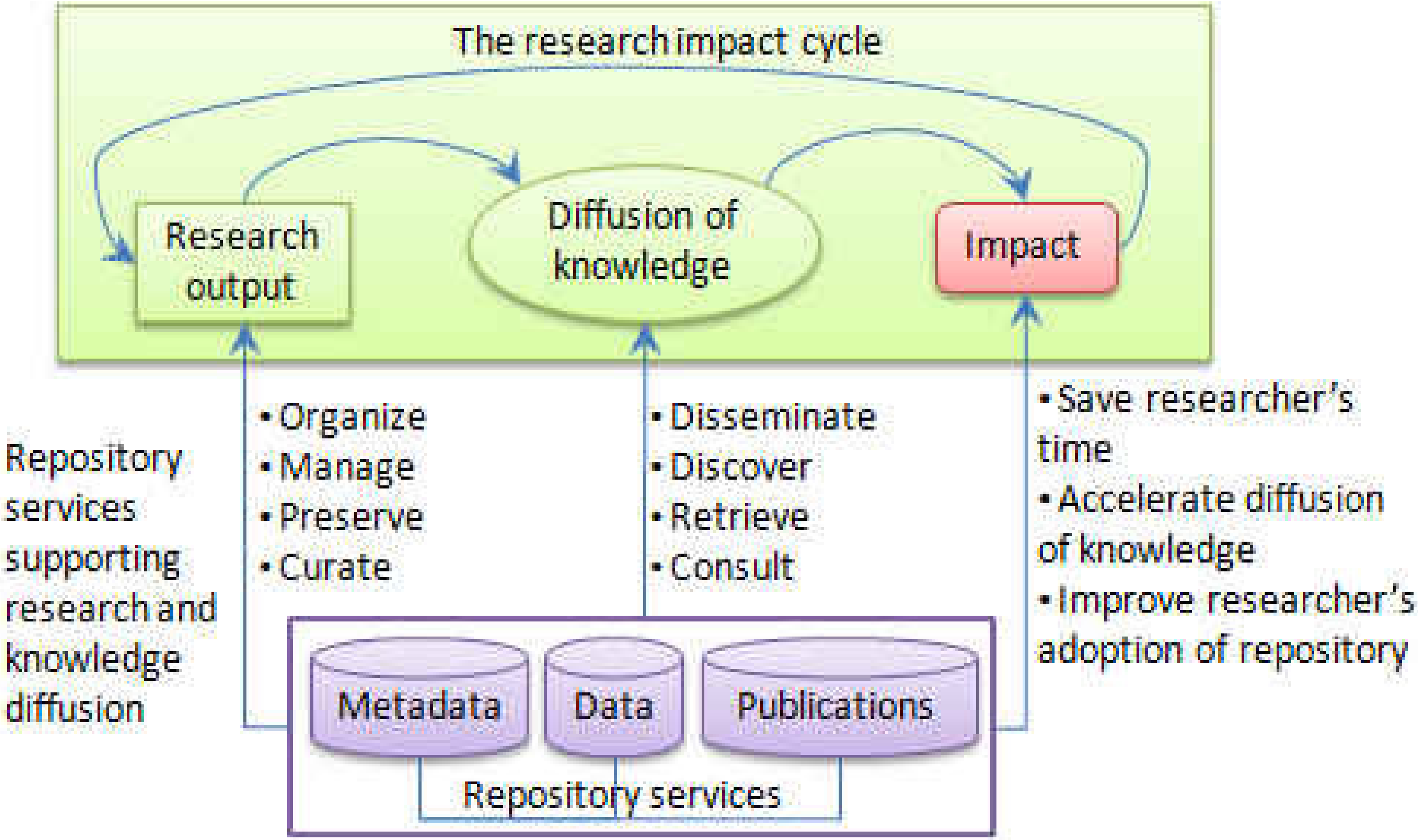
TADS is designed to compare the effectiveness of established treatments for teenagers suffering from major depressive disorder (MDD). The treatments are: psychotherapy ("talking therapy"); medication; and the combination of psychotherapy and medication. Altogether, 432 teenagers (both males and females) ages 12 to 17, will take part in this study at 12 sites in the United States.

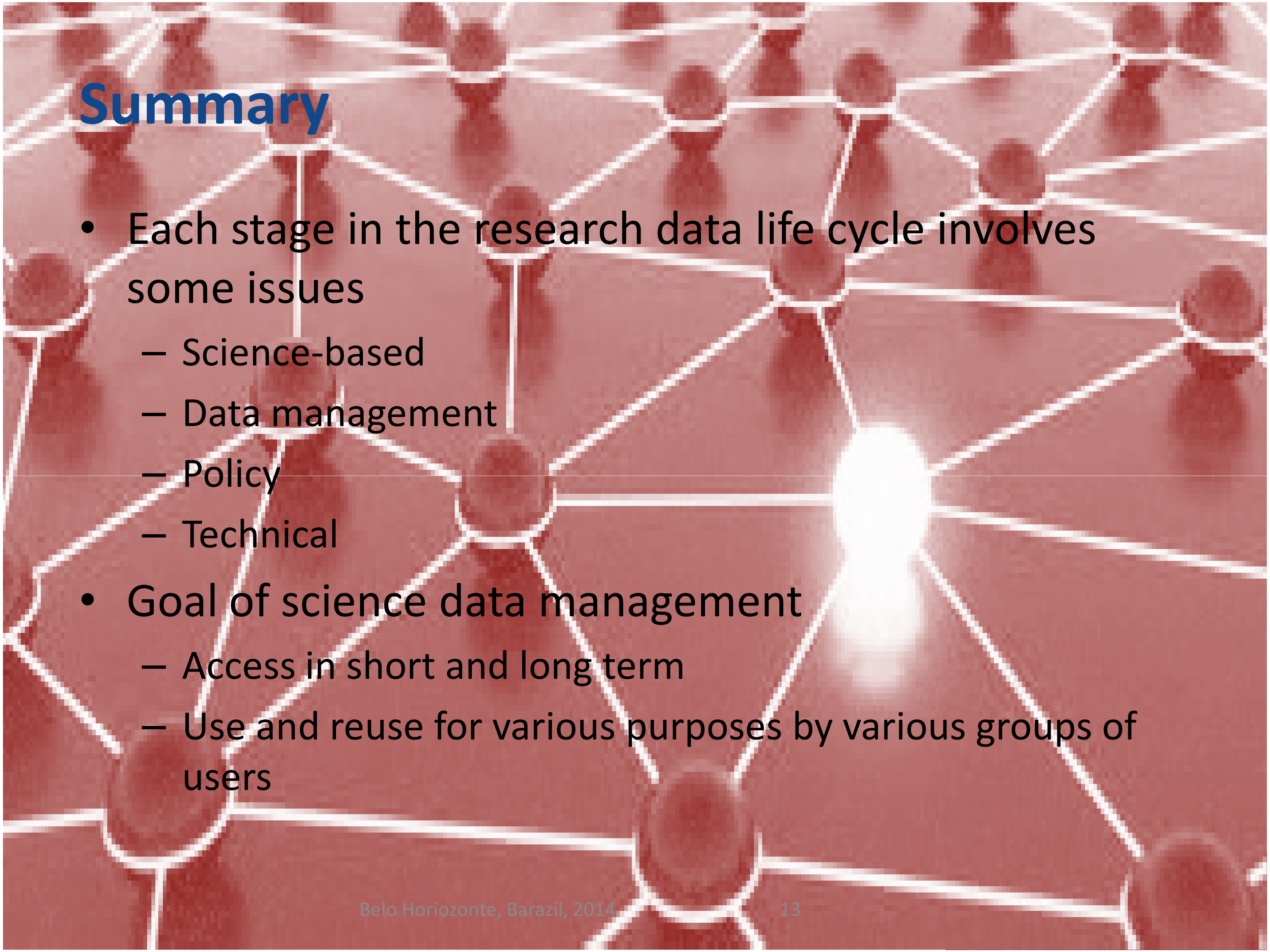
The **TADS** design will provide answers to the following questions: What is the long-term effectiveness of medication treatment of teenagers who have major depression? What is the long-term effectiveness of a specific psychotherapy ("talking therapy) in the treatment of teenagers who have major depression? How does medication treatment compare with psychotherapy in terms of effectiveness, tolerability and teenager and family acceptance? And, What is the cost-effectiveness of medication, psychotherapy and combined treatments?

The medication being used in this study is called fluoxetine. Fluoxetine is also known as Prozac. Research has shown that medications like Prozac help depression in young persons. Fluoxetine has been approved by the FDA for use in the treatment of child and adolescent (ages 7 to 17 years) depression.

The psychotherapy or "talking therapy" being used in this study is called Cognitive Behavioral Therapy (CBT). CBT is a talking therapy that will teach both the teenager and his or her family member (e.g., parent) new skills to cope better with depression. Specific topics include education about depression and the causes of depression, setting goals, monitoring mood, increasing pleasant activities, social problem-solving, correcting negative thinking, negotiation, compromise and assertiveness. CBT sessions may also help with resolving disagreements as they affect families.

RDM, research impact, and value





Summary

- Each stage in the research data life cycle involves some issues
  - Science-based
  - Data management
  - Policy
  - Technical
- Goal of science data management
  - Access in short and long term
  - Use and reuse for various purposes by various groups of users

Research Data Management as Infrastructure Services

What is an infrastructure?

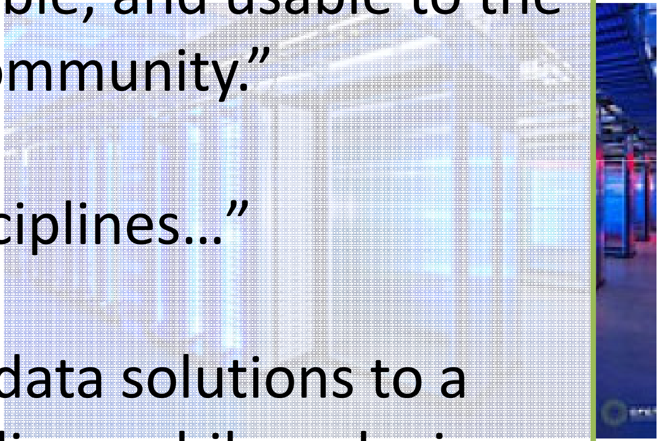
The underlying foundation or basic framework (as of a system or organization).

The system of public works of a country, state, or region.

The resources (as personnel, buildings, or equipment) required for an activity.

http://www.merriam-webster.com/dictionary/infrastructure

What is data infrastructure?



"a sustainable data infrastructure that will be discoverable, searchable, accessible, and usable to the entire research and education community."

"usable by multiple scientific disciplines…"

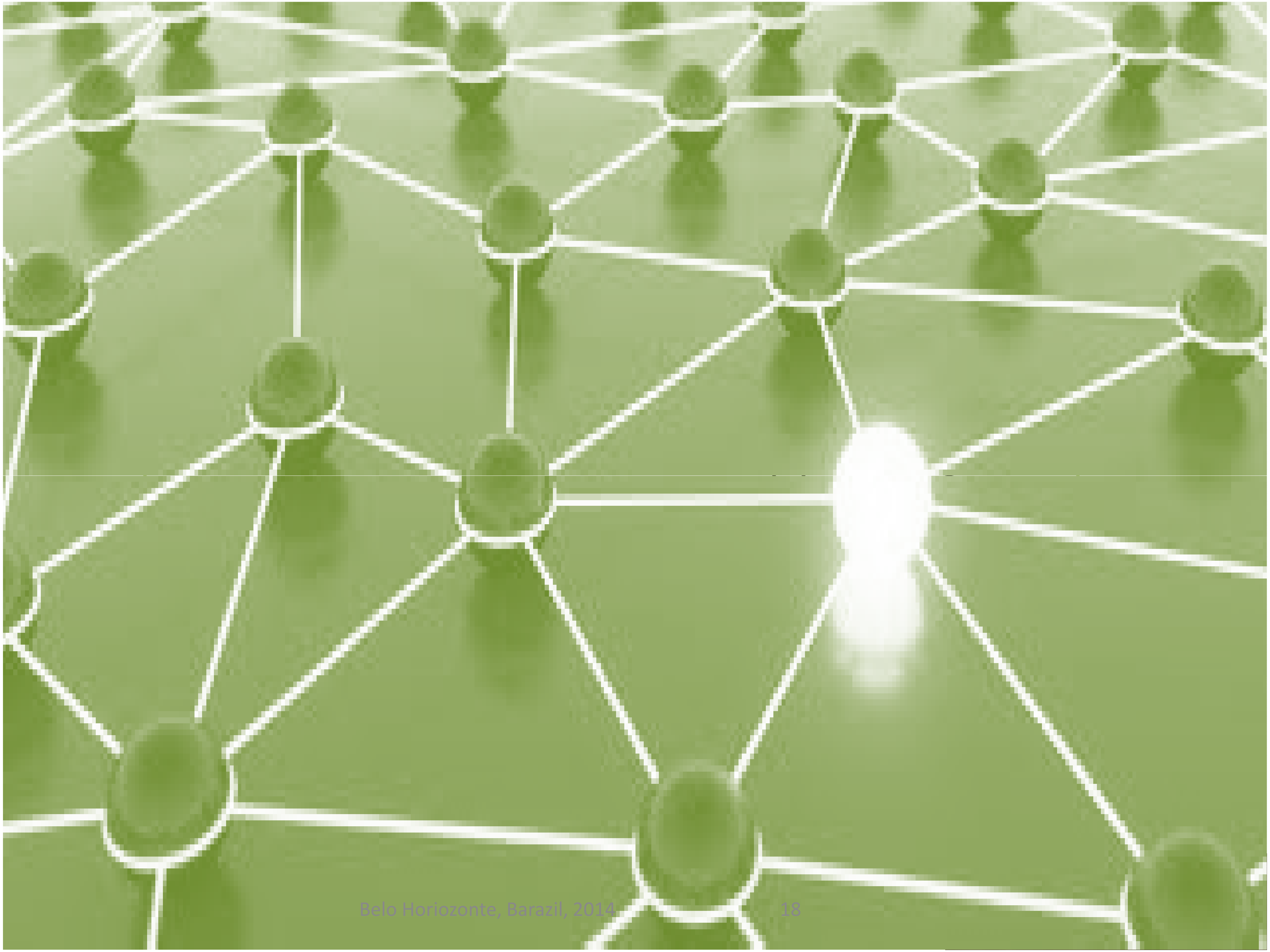
"…that can support and provide data solutions to a broader range of scientific disciplines while reducing duplicative efforts."

http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504776

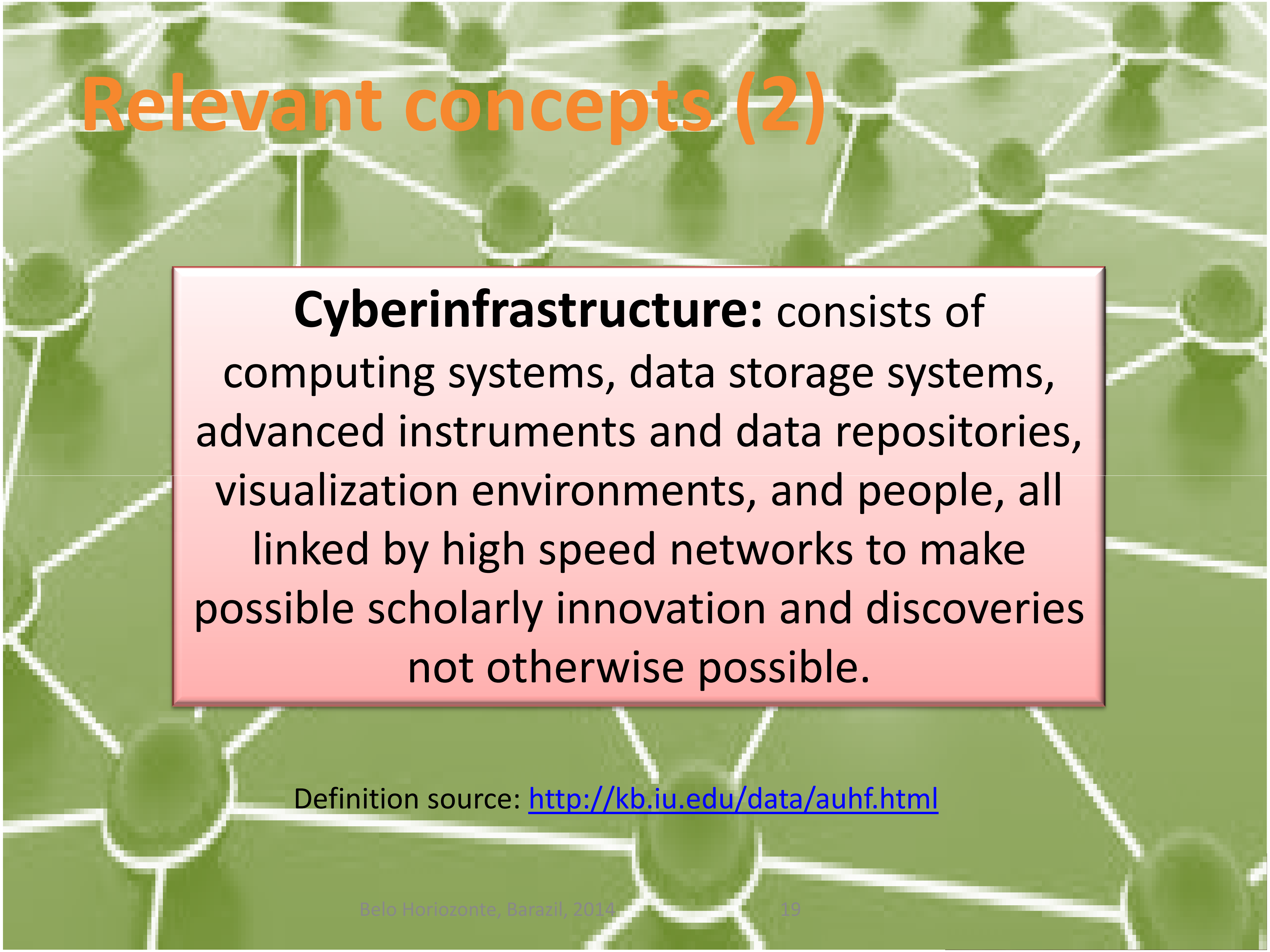
Nature of an infrastructure

- **Embeddedness**. Infrastructure is sunk into, inside of, other structures, social arrangements, and technologies.
- **Transparency**. Infrastructure does not have to be reinvented each time of assembled for each task, but invisibly supports those tasks.
- **Reach or scope beyond a single event or a local practice**.
- **Learned as part of membership**.
- **Links with conventions of practice**.
- **Embodiment of standards**.
- **Built on an installed base**.
- **Becomes visible upon breakdown.**
- **Is fixed in modular increments**, not all at once or globally.

(Star & Ruhleder, 1996)



Relevant concepts (2)



Cyberinfrastructure: consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked by high speed networks to make possible scholarly innovation and discoveries not otherwise possible.

Definition source: http://kb.iu.edu/data/auhf.html

Relevant concepts (3)

Collections: information items brought together for some specific purpose or with at least one feature in common. They may be

- generated by an institution or project
- gathered for a discipline or by an individual

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf

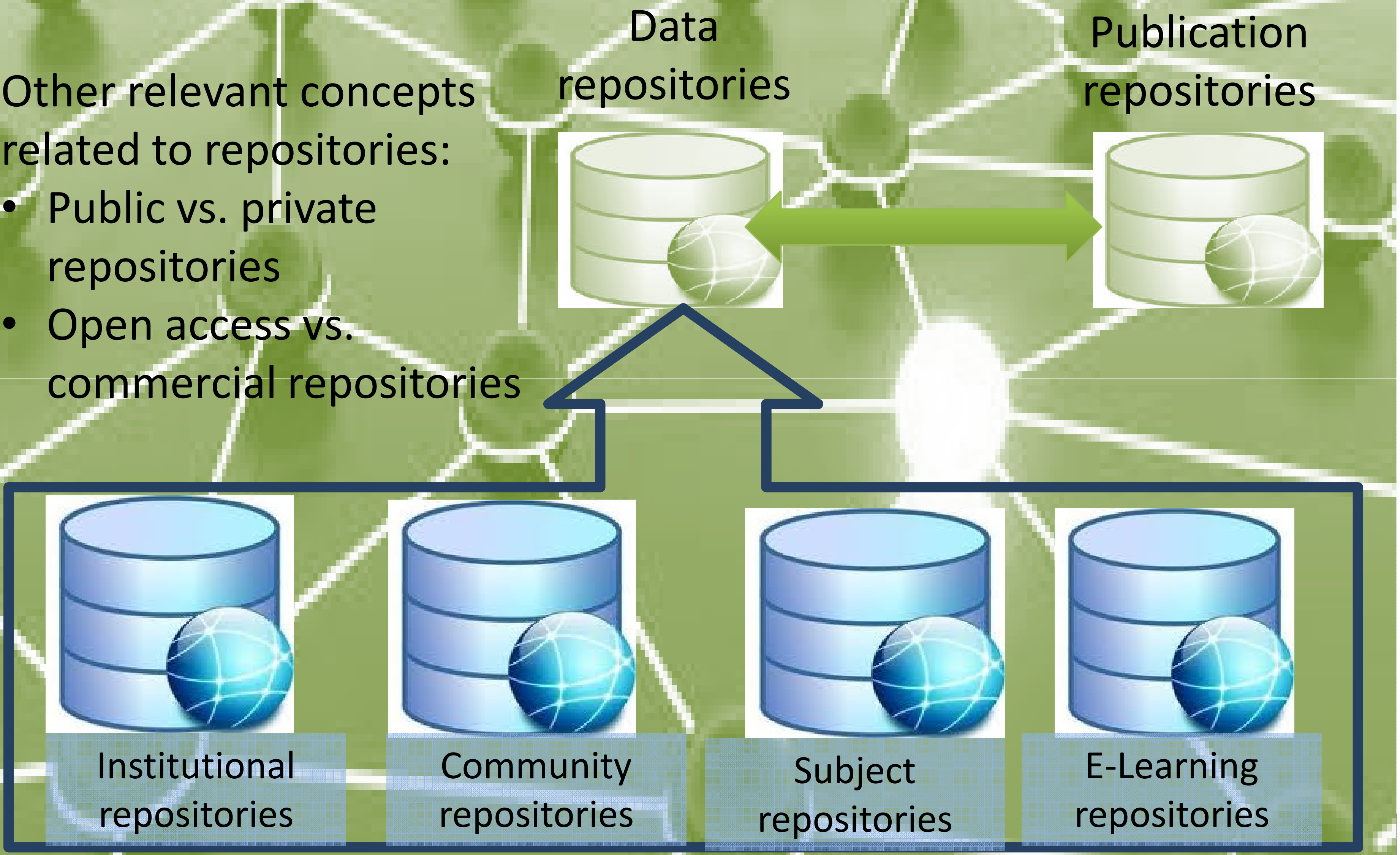
Repositories: the constructs that hold collections and facilitate their use.

- Narrowly, they mean storage equipment and supporting computer programs
- Wider definition: they include the management framework, services, and tools associated with a repository as well as the storage machinery itself.

Relevant concepts (3, cont'd)

Other relevant concepts related to repositories:

- Public vs. private repositories
- Open access vs. commercial repositories



Data repositories

Publication repositories

Institutional repositories

Community repositories

Subject repositories

E-Learning repositories

Repositories as an important part of data infrastructure services

- Content is deposited in the repository, whether by content creator, owner, or third party
- The repository architecture manages content as well as metadata
- The repository offers a minimum set of basic services such as put, get, search, and access control
- The repository must be sustainable and trusted, well-supported and well-managed

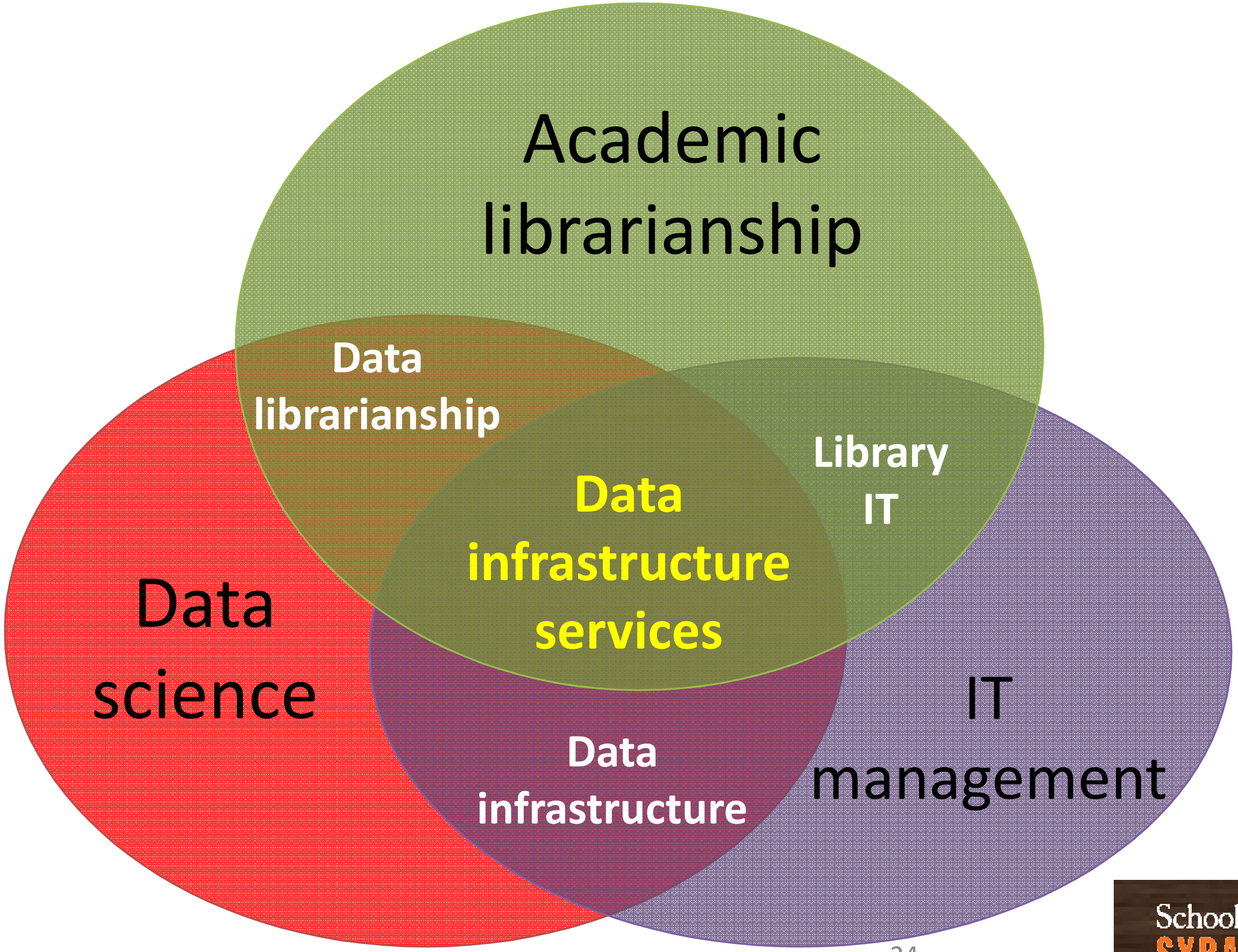
http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf

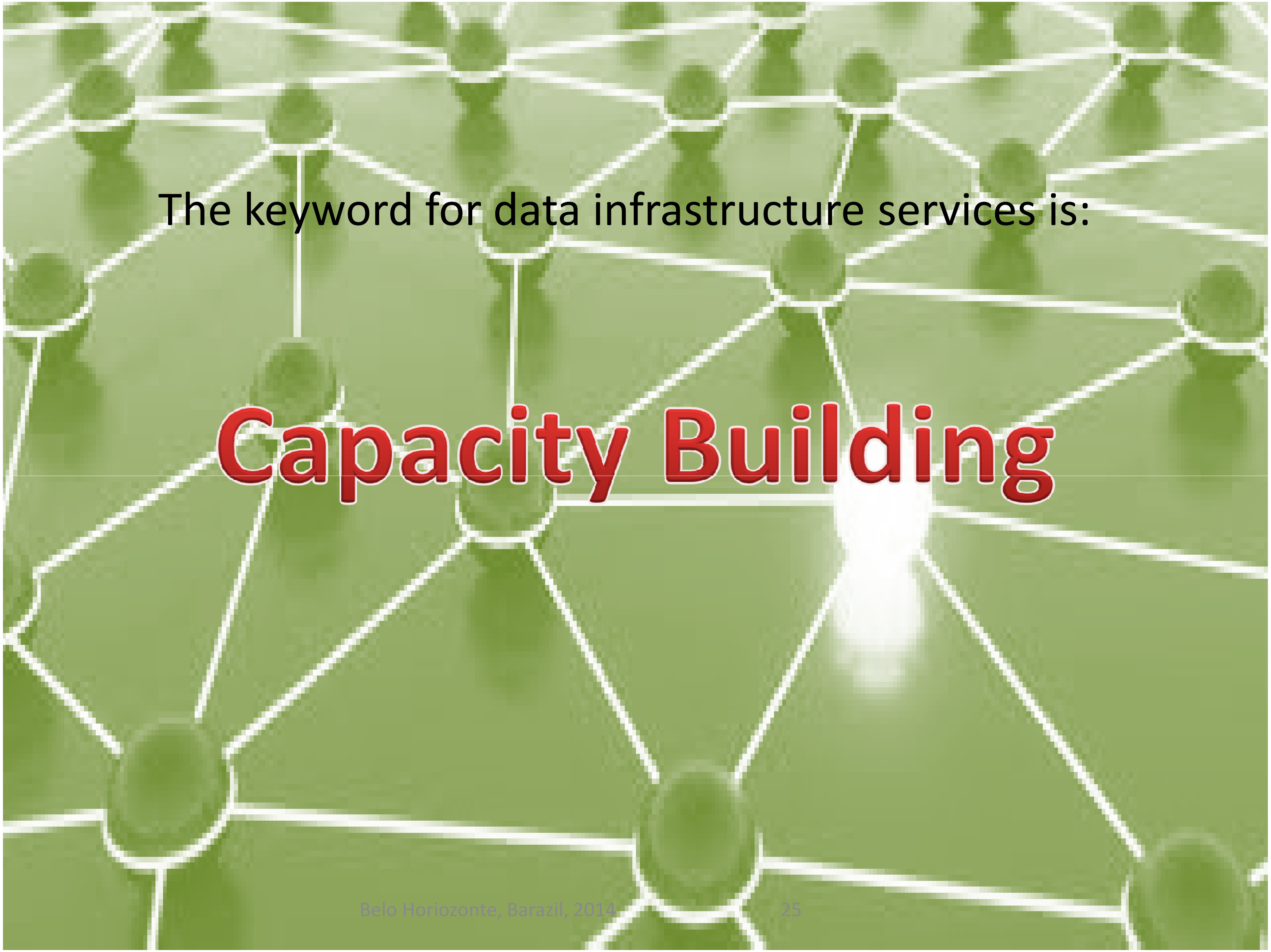
Data science: an emerging area of work

“An emerging area of work concerned with the collection, presentation, analysis, visualization, management, and preservation of large collections of information.”

(Stanton, 2012)

Data infrastructure services and academic libraries

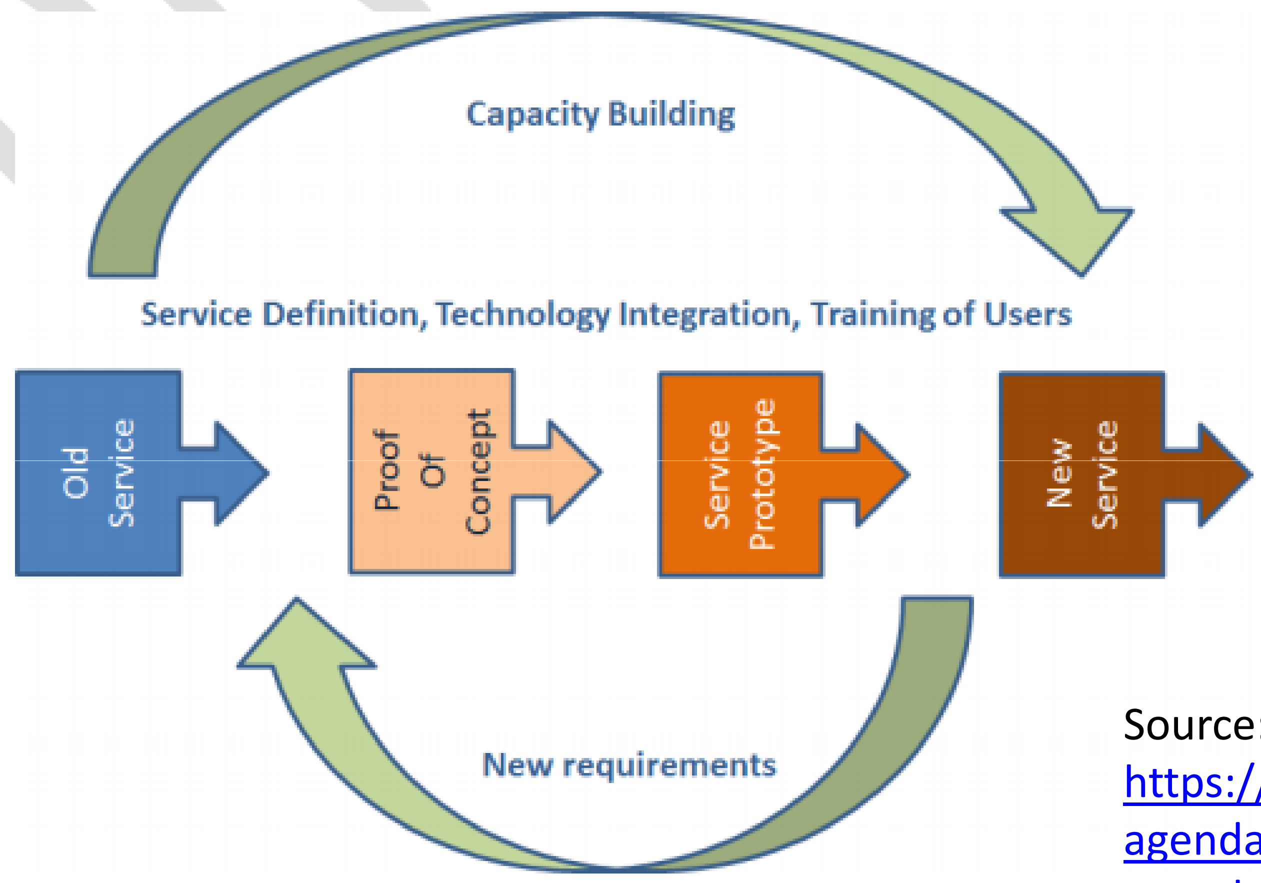




The keyword for data infrastructure services is:

Capacity Building

Capacity building lifecycle

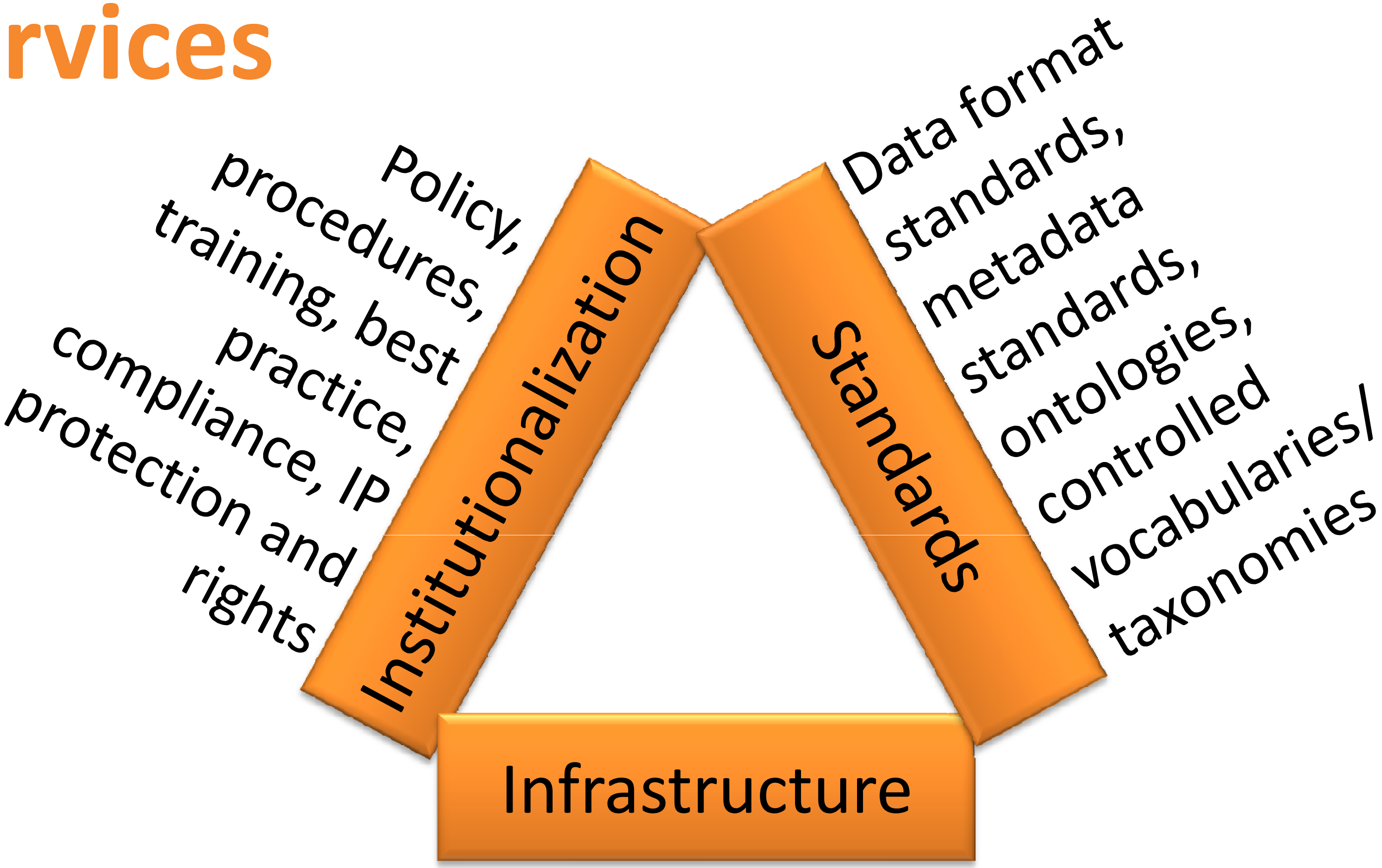


capacity building lifecycle

Source: https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/data_einfra_h2020_fiches_on-line_consult.pdf , p. 9

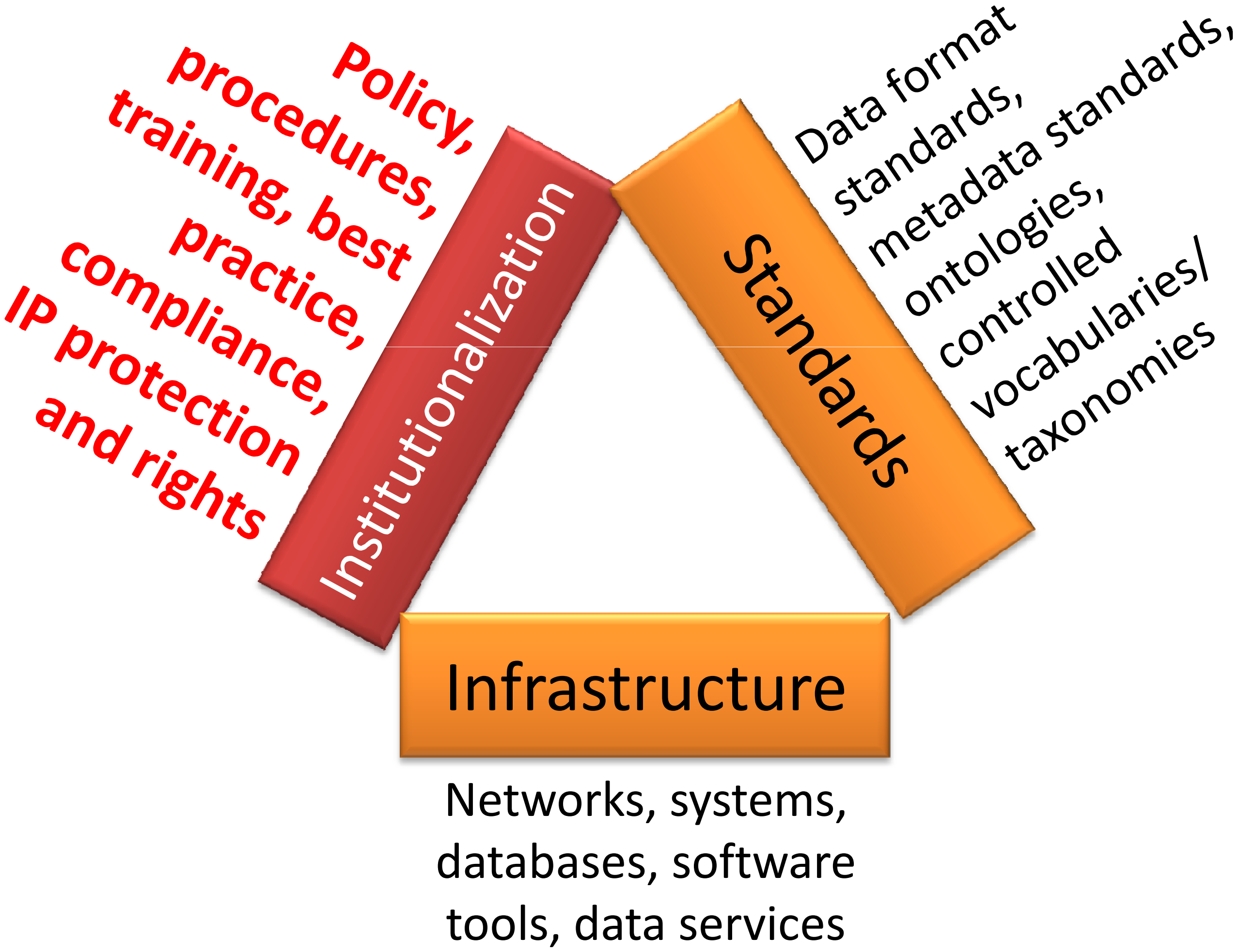
Challenges and key issues in the data-driven research environment

Three pillars in data infrastructure services


Policy, procedures, training, best practice, compliance, IP protection and rights
Institutionalization
Standards
Data format standards, metadata standards, ontologies, controlled vocabularies/ taxonomies
Infrastructure

Networks, systems, databases, software tools, data services

What is institutionalization?
Why do you need institutionalize research data management?
How can you institutionalize RDM?



Policy, procedures, training, best practice, compliance, IP protection and rights

Institutionalization

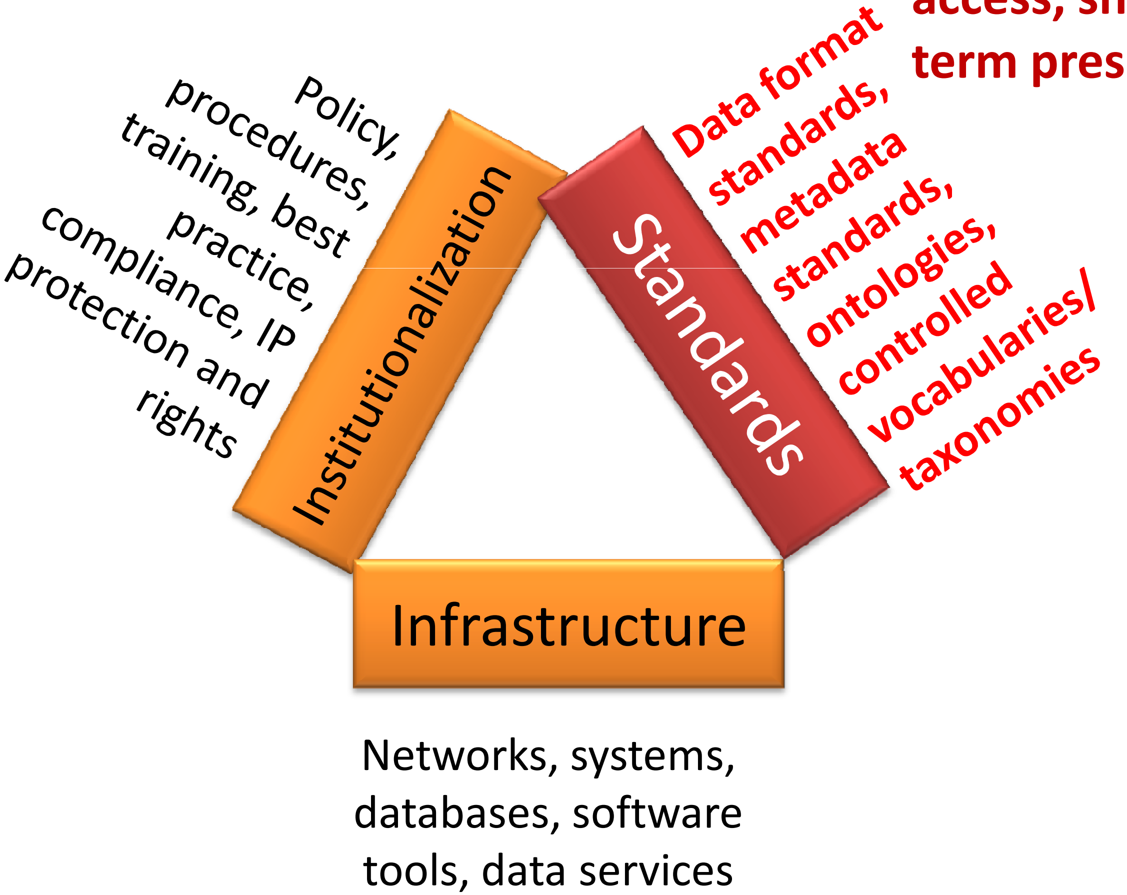
Standards

Data format standards, metadata standards, ontologies, controlled vocabularies/ taxonomies

Infrastructure

Networks, systems, databases, software tools, data services

How much do you know about data and metadata?
How does the nature of data affect metadata?
How does metadata affect data access, sharing, reuse, and long-term preservation?



Institutionalization

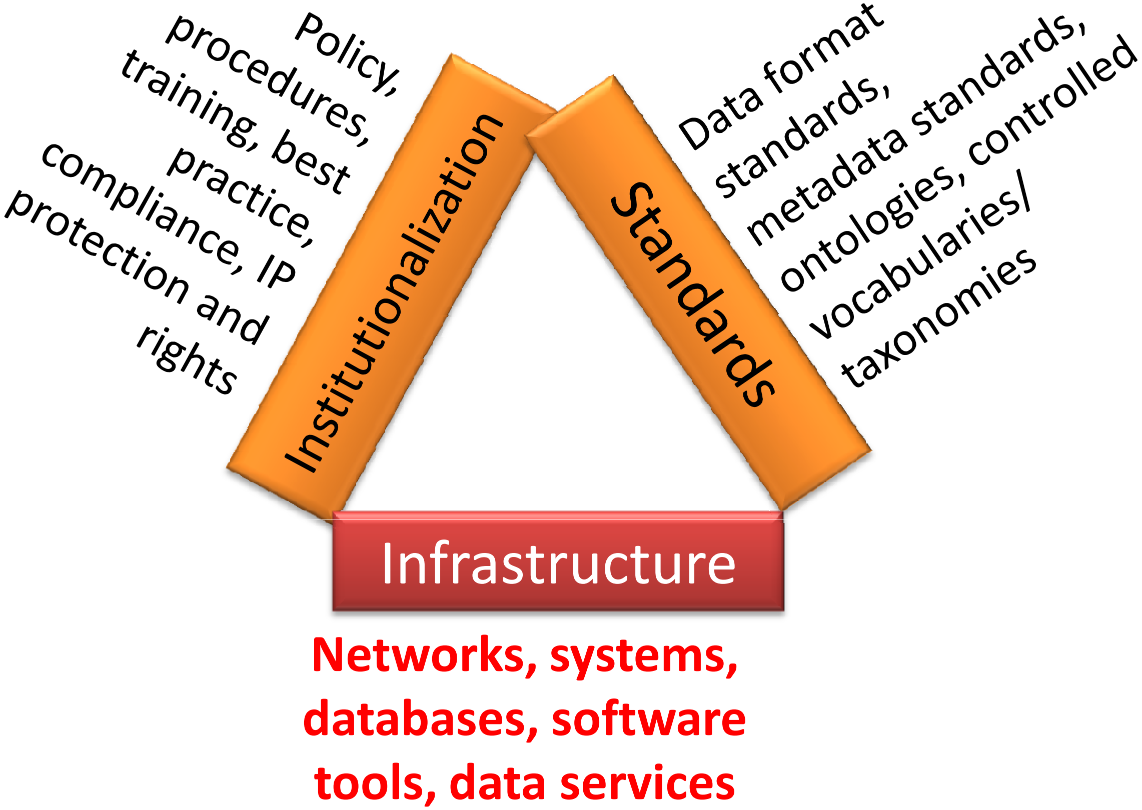
Policy, procedures, training, best practice, compliance, IP protection and rights

Standards

Data format standards, metadata standards, ontologies, controlled vocabularies/ taxonomies

Infrastructure

Networks, systems, databases, software tools, data services



Policy, procedures, training, best practice, compliance, IP protection and rights

Institutionalization

Standards

Data format standards, metadata standards, controlled ontologies, vocabularies/ taxonomies

Infrastructure

Networks, systems, databases, software tools, data services

What is data infrastructure and Data infrastructure services?
Why do you need to build a data infrastructure?
What is the key in building a data infrastructure?

What lies before research libraries?

Challenges

- Extremely rapid growth and sheer volume of data and metadata, in and out of repositories
- Developing and maintaining standards, database schemas, ontologies, etc. to improve interoperability
- Initiatives started to create global information infrastructures at semantic level
- Difficulties relating to data deposit, both technically and behaviorally
- Repositories are under-staffed

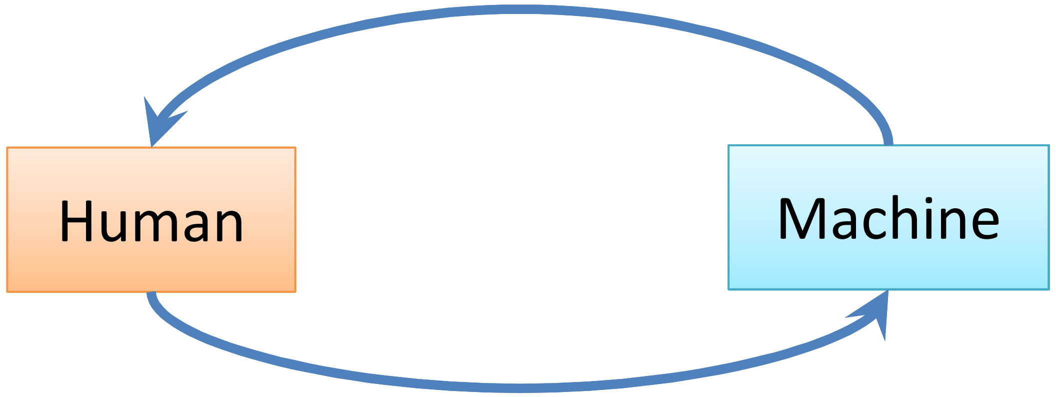
Where are the gaps?

Vision

- A reliable, sustainable data infrastructure and services for data repositories
- A high quality information space
- Information that is readily available
- A well-managed, accountable repository infrastructure

Understanding data services

For whom?



Infrastructure type of services:

- National
- Institutional

Data services have evolved…

ARL survey report 2010

Finding relevant data 83%

Developing data management plans 79%

Finding and using available technology infrastructure and tools 76%

Developing tools to assist researchers 76%

Archiving and curating relevant data and curating it for long-term preservation and integration across datasets

Providing curatorial and data Stewardship services

Raising awareness and user training

(Soehner, Steeves, & Ward, 2010)

ARL Survey report 2013 (N=72)	
Providing an institutional repository	89%
Locating & using existing data sources	94%
GIS and geospatial analysis, support	85%
Dataset purchase, acquisition, subscriptions	81%
Copyright & patent advising	75%
General statistical software support	58%
Data visualization support	36%
Data analysis support	39%
Data mining	28%
Database design & management	28%
Programing/software development	24%
Other data support services	29%

(Fearon et al., 2013)

Data Management Plan (DMP) services

- Online DMP services:
  - Explanation of DMP requirements by different funding agencies and/or NSF directorates
  - Guidelines for creating DMPs
  - Template examples of DMPs
  - A tool or resource for DMP creation
  - A data planning checklist
  - Copyright considerations, data citation guidelines, metadata examples, info about digital repository services
  - …

What data services? (2)

- Submission of data
- Data export
- Data format conversion /transformation
- Access to data (discovering and obtaining data)
- IP protection and management
- Educational offerings
- Technical assistance including data management and manipulation services
    - Access to computing facilities
    - Curation
    - Archive and preservation tools
    - Information
- Print and publication services
- Marketing
- Publicity
- Software development services

Source: Marcial & Hemminger, 2010

Key issue 1: Staffing

- ARL survey report (2013):
  - Subject librarian or liaison (50%)
  - Digital (38%)
  - Data librarian (18%)
  - Metadata (17%)
  - Data services (13%)
  - GIS or Geospatial (12%)
  - Research data (11%)
  - Curation (11%)
  - Repository (10%)
  - Software or systems (9%)
  - Data management (9%)

Key issue 2: funding

- Internal library regular budget 98%
- Direct administrative funding (separate from library funds) 11%
- External grant funding 11%
- Internal library temporary or special project budget 9%
- Department or research institute/project group funds 6%
- Endowment fund 6%
- Fee to researcher or researcher’s grant 4%
- Facilities and administrative (F&A) funding 2%
- Other source of funding 9%

(N=57, from Fearon et al., 2013)

Key issue 3: training

- Training / Experience most important to RDM:
  - Subject domain expertise
  - Digital/data curation training
  - IT technology or services experience
  - Library MLS/MLIS training

- Particularly important:
  - Research methods and data analysis
  - Research data management
  - Data curation
  - Scholarly communication
- Other skills and training:
  - Identifying and applying metadata standards
  - Digital preservation
  - Data ownership policies
  - Ethical and legal issues
  - Data security
  - Data sharing and access
  - Data storage and backup planning
  - Data retention policy
  - Data citation

Key issue 4: data policies

- Institutional data policies
- Project data policies

Data policies are a major component of the institutionalization of research data management

How can you make a contribution? (1)

- Capabilities of the 21st century academic library information professional
  - Deep Subject, Process, or Technical Expertise
  - Deep Service Commitment
  - Commitment to Research and Development
  - Commitment to Assessment and Evaluation
  - Communication and Marketing Skills
  - Project Development and Management Skills
  - Political Engagement
  - Resource Development Skills
  - Commitment to Rigor
  - Entrepreneurial Spirit
  - Commitment to Collaboration
  - Leadership/Inspirational Capacity

http://www.arl.org/storage/documents/publications/2012-hrsym-pres-neal-j.pdf

How can you make a contribution? (2)

“Academic library staff must integrate and mainstream digital library services, digital archiving and preservation, repository development, digital publishing, and instructional technologies into the core of library budgeting, staffing and organization.”

“Academic library staff must be embedded in the e-research cyberinfrastructure and scholarly communication processes, and be integral to the systems of research information management.”

- Support the needs of data-driven research
  - Federal/funding agency
  - Massive data sets
  - Unstructured data/curation
  - Extraction
  - Distribution
  - Collaboration
  - Visualization
  - Simulation
  - Preservation

http://www.arl.org/storage/documents/publications/2012-hrsym-pres-neal-j.pdf

Concluding remarks

- Data management and services are a new territory and require new thinking of research library’s roles
- There are ample opportunities for research librarians to make a contribution to capacity building
- Learn the lessons from earlier initiatives
- Transforming the library image needs to transform the library tradition first

Thank you!

enancib
2014
Além das "nuvens":
expandindo as fronteiras
da Ciência da Informação

References

- Fearon, D. Jr., Gunia, B., Pralle, B. E., Lake, S., & Sallans, A. L. (2013). Research Data Management Services. Washington, DC: Research Library Association.
- Star, S.L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information space. Information Systems Research, 7(1): 111-134.
- Marcial, L. H. & Hemminger, B. M. (2010). Scientific Data Repositories on the Web: An Initial Survey. *Journal of the American Society for Information Science and Technology*, 61(10): 2029-2048.
- McCormick, T. (2009). A Web services taxonomy: not all about the data. http://tjm.org/public/Web-Services-Taxonomy_McCormick_v1.1.pdf
- Venugopal, S., Buyya, R., & Ramamohanarao, K. (2006). A taxonomy of data grids for distributed data sharing, management, and processing. *ACM Computing Surveys*, 38(1): http://arxiv.org/pdf/cs.DC/0506034
- Soehner, C., Steeves, C., & Ward, J. (2010). E-Science and data support services: A study of ARL member institutions. http://www.arl.org/bm~doc/escience_report2010.pdf
- Stanton, J. (2012). Introduction to Data Science. http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

Examples of (open) data infrastructure services

- The Institute for Quantitative Social Science repository: http://www.iq.harvard.edu/
- Inter-University Consortium for Political and Social Research (ICPSR): http://www.icpsr.umich.edu/icpsrweb/landing.jsp
- The Dryad Digital Repository: http://datadryad.org/
- Data Observation Network for Earth: http://www.dataone.org/
- Datalib: http://databib.org/ (a registry/directory/catalog of research data repositories)
- Registry of Research Data Repositories: http://www.re3data.org/