Open Science, FAIR data and Cyberinfrastructure

Beth Plale
Professor, Indiana University Bloomington
Science Advisor for Public Access, US National Science Foundation

Open science is good science

NSF'S PUBLIC ACCESS PLAN:

Today's Data, Tomorrow's Discoveries

Increasing Access to the Results of Research Funded by the National Science Foundation

National Science Foundation

March 18, 2015

nature

Nature Research and Springer Formally Endorse Guidelines that Promote Transparency in the Process of Science

March 16, 2017 | The Center for Open Science is pleased to announce that Nature Research and Springer journals have today become signatories of the Transparency and Openness Promotion (TOP) Guidelines.

NEWS | 23 November 2018 | Vienna, Austria | Research and Innovation

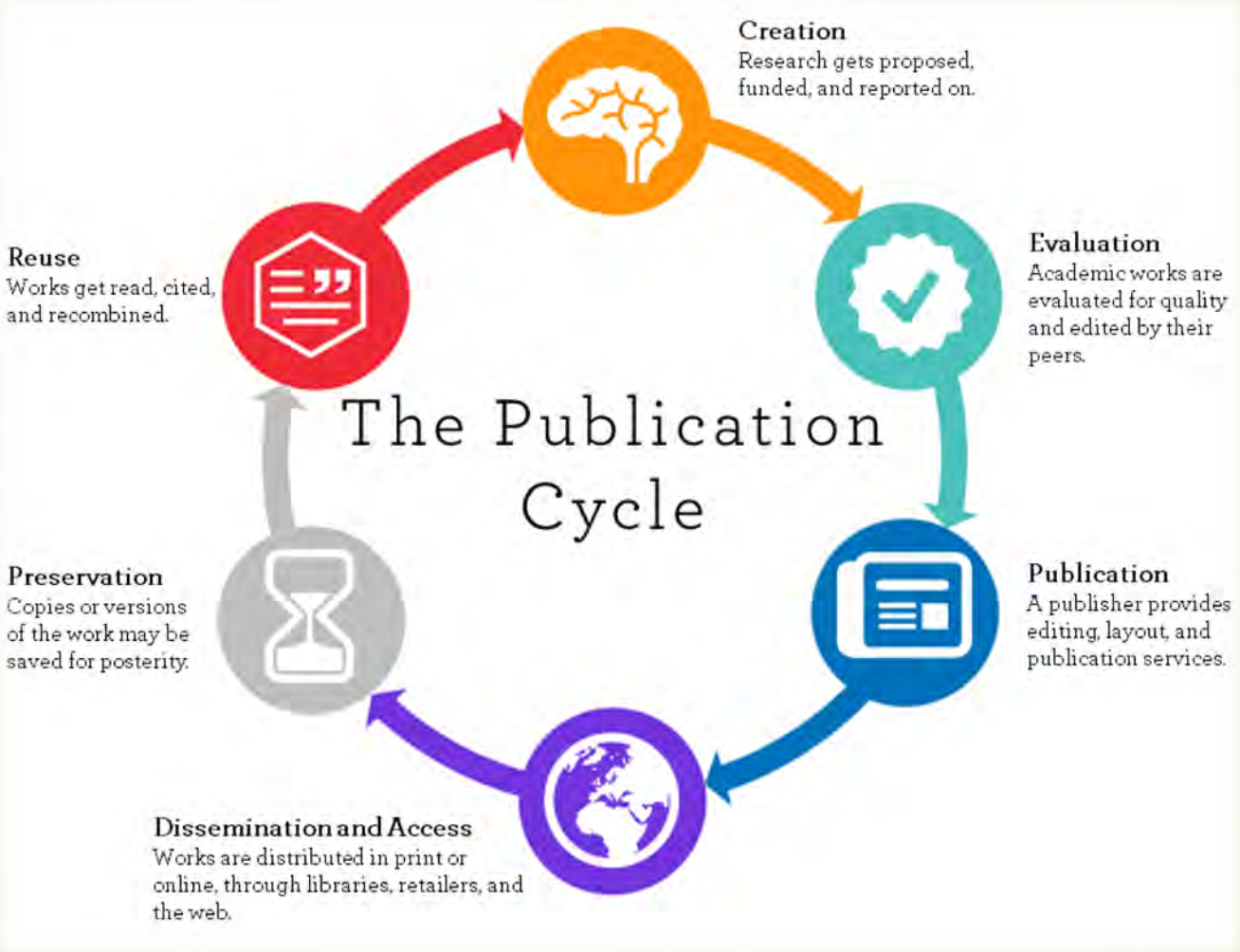
Commission launches European Open Science Cloud

Following a major effort by the European Commission, the Member States and the scientific community, the European Open Science Cloud (EOSC) was launched today to provide a safe environment for researchers to store, analyse and re-use data for research, innovation and educational purposes. The Commission presented the governance structure and the portal to EU science ministers and future users at an Austrian EU Presidency conference in Vienna.

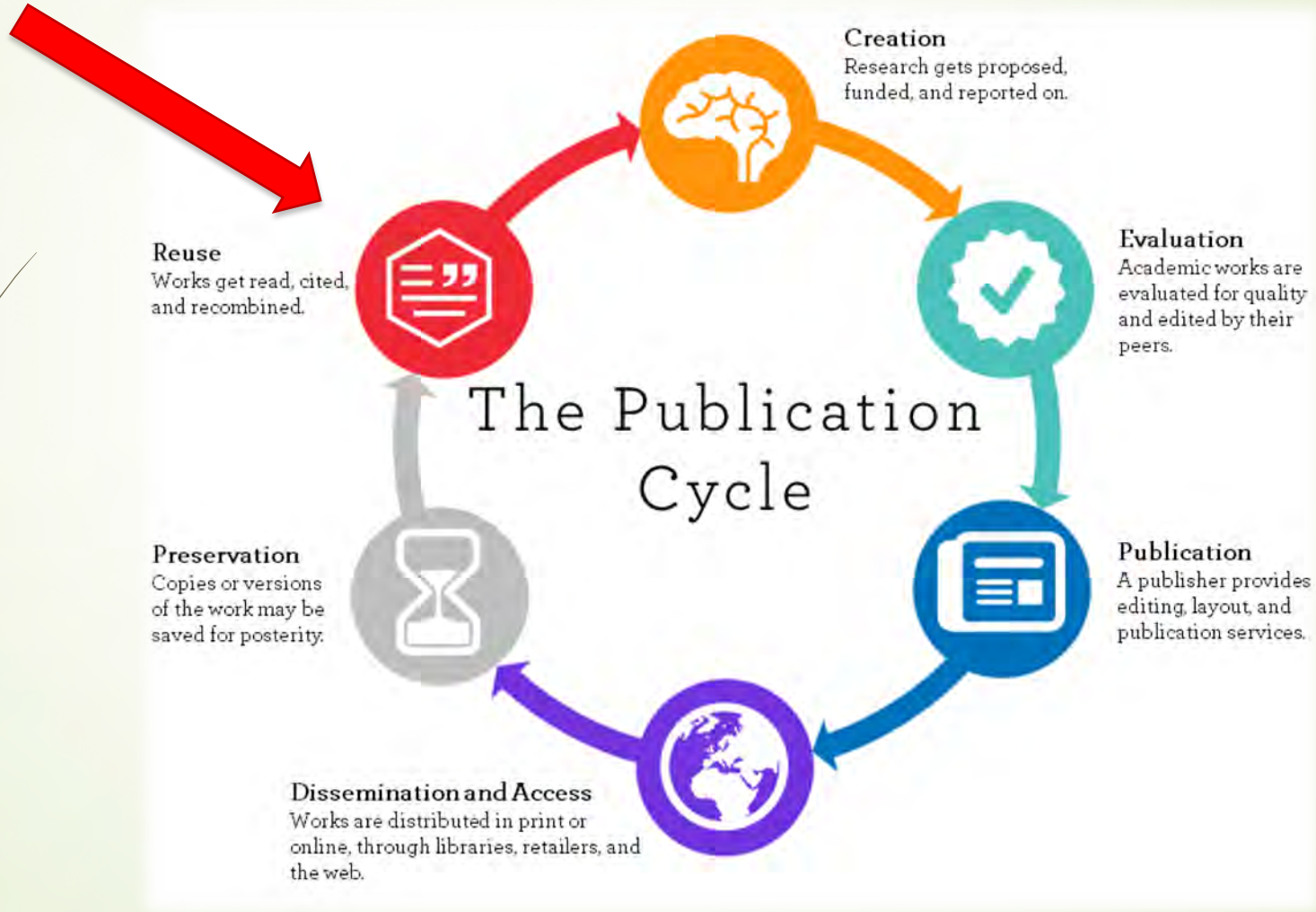
Do social science research findings published in Nature and Science replicate?

Aug. 27, 2018 | Replications of 21 high-profile social science findings demonstrate challenges for reproducibility and suggest solutions to improve research credibility.

Open science predicated on value of data created through research



Open science predicated on value of data created through research



Federal Action in Open Science

Investigators are expected to *share with other researchers*, at no more than incremental cost and within a reasonable time, the *primary data*, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants.

US National Science Foundation

Data Management Plans (DMP)

- Researcher writes a *Data Management Plan* for the important data that they expect to create during course of their research
- By National Science Foundation: "What constitutes reasonable data management and access will be determined by the community of interest *through process of peer review and program management.*" [Data Management & Sharing Frequently Asked Questions, National Science Foundation]

European Commission

Open Research Data pilot in EU Horizon 2020

Participating projects are required to develop a Data Management Plan, in which they will specify what data will be open. In previous work programmes, the ORD Pilot was limited to some specific areas of Horizon 2020. Starting with the 2017 work programme, however, the ORD pilot was extended to cover **all thematic areas** of Horizon 2020, thus realising the Commission's ambition of "open research data per default" (but allowing for opt-outs).

Why enable data reuse?

- Encourages scientific enquiry and debate:
  - *Encourages improvement* and validation of research methods
  - *Maximizes transparency* and accountability through scrutiny of research findings.
- Promotes innovation and potential new data uses:
  - leading to new collaborations between data users and data creators
  - *Reducing cost* of duplicating data collection
  - *Increasing impact* and visibility of research
  - *Providing credit* to the researcher as a research output in its own right

ands
AUSTRALIAN NATIONAL DATA SERVICE

UK Data Archive

How make data available for reuse

Good data management is the key for data (re)use:

- Planning for reuse and publication from the start.
- Recognition of others' data through appropriate citation.
- Appropriate rules of use through simple and explicit data licensing approaches.
- Sufficient metadata describing how the data has been specified, collected, analyzed and transformed.
- Use of standard vocabularies in the metadata also enables reuse.
- Data resulting from research needing ethical permission and oversight needs particular preparation if it is to be shared.

The most effective way to get your data reused is to publish it.

ands
AUSTRALIAN NATIONAL DATA SERVICE

UK Data Archive

Four (personal) observations within open science

1. Data valuation
2. FAIR principles
3. Open as possible
4. Cyberinfrastructure role in reproducibility

Not all research data created in the context of science is data of value

Data value: value of data (object, product, or collection) to science and society either as part of larger scholarly record (inherent value) or through enabling new discoveries

More data is generated in course of science than can be kept.

- *What data can be thrown away?*
- *How long should a dataset be kept?*
- *Who decides?*

FAIR principles

- A concise and measurable set of principles for scientific data management
- Developed in 2015 under umbrella of Force11
- Data objects are
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable

FORCE11

The Future of Research Communications and e-Scholarship

FAIR Guiding Principles

To be Findable:

F1. Data are assigned a globally unique and eternally persistent identifier (PID)

F2. Data are described with rich metadata

To be Accessible:

A1. Data are retrievable by their identifier using a standardized communications protocol

A2. Metadata are accessible, even when data are no longer available

FORCE11
The Future of Research Communications and e-Scholarship

FAIR Guiding Principles

To be Interoperable

I.1. Data is machine-actionable

I.2. Data formats utilize shared vocabularies and/or ontologies

To be Re-usable

R.2 Data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources

plale@indiana.edu

FORCE11
The Future of Research Communications and e-Scholarship

The FAIR Data Principles set out requirements for data to be processed in an automated way

WORLD ECONOMIC FORUM
COMMITTED TO IMPROVING THE STATE OF THE WORLD

Findable:
"Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets"
- e.g. Able to locate data by individual patient, patient segment, intervention, outcome metric

Accessible:
"Stored for long term such that they can be easily accessed and / or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content"
- e.g. Patients should be able to access parts of their own data via a patient controlled record

Interoperable:
"Ready to be combined with other datasets by humans as well as computer systems"
- Semantic interoperability: mapped data taxonomies across diseases and population groups e.g. consistent methodology & scale for measuring pain / quality of life
- Technical interoperability: specifications to allow different systems to communicate with each other

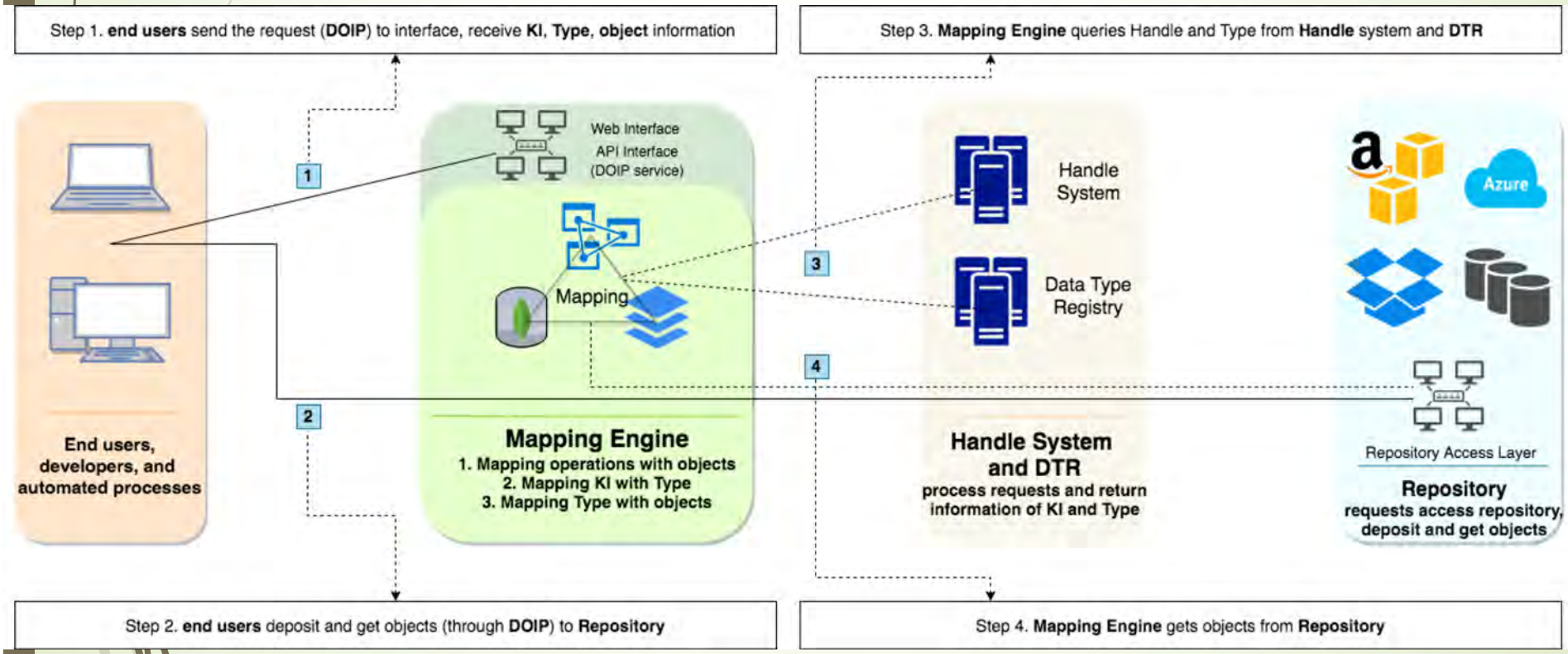
Reusable:
"Ready to be used for future research and to be processed further using computational methods"
- e.g. Outcomes data should be available for the long-term for systematic analysis or clinical research (with permission from data owner)

Important that interoperable datasets can be interpreted by computer systems: to (semi) automatically combine different data sources for richer knowledge discovery

Source: Dutch Techcentre for Life Sciences

eRPID Testbed

- Available for community use
- Assigns and resolves Handles
- Subdomain will be set up for project; Handles are test handles
- Easy to convert to DOIs once experimentation stage over



- See Yu Luo for more details

The books I want to text mine are under copyright

Video from my sensors captures everyday life of citizens

Restricted Data and Open Science

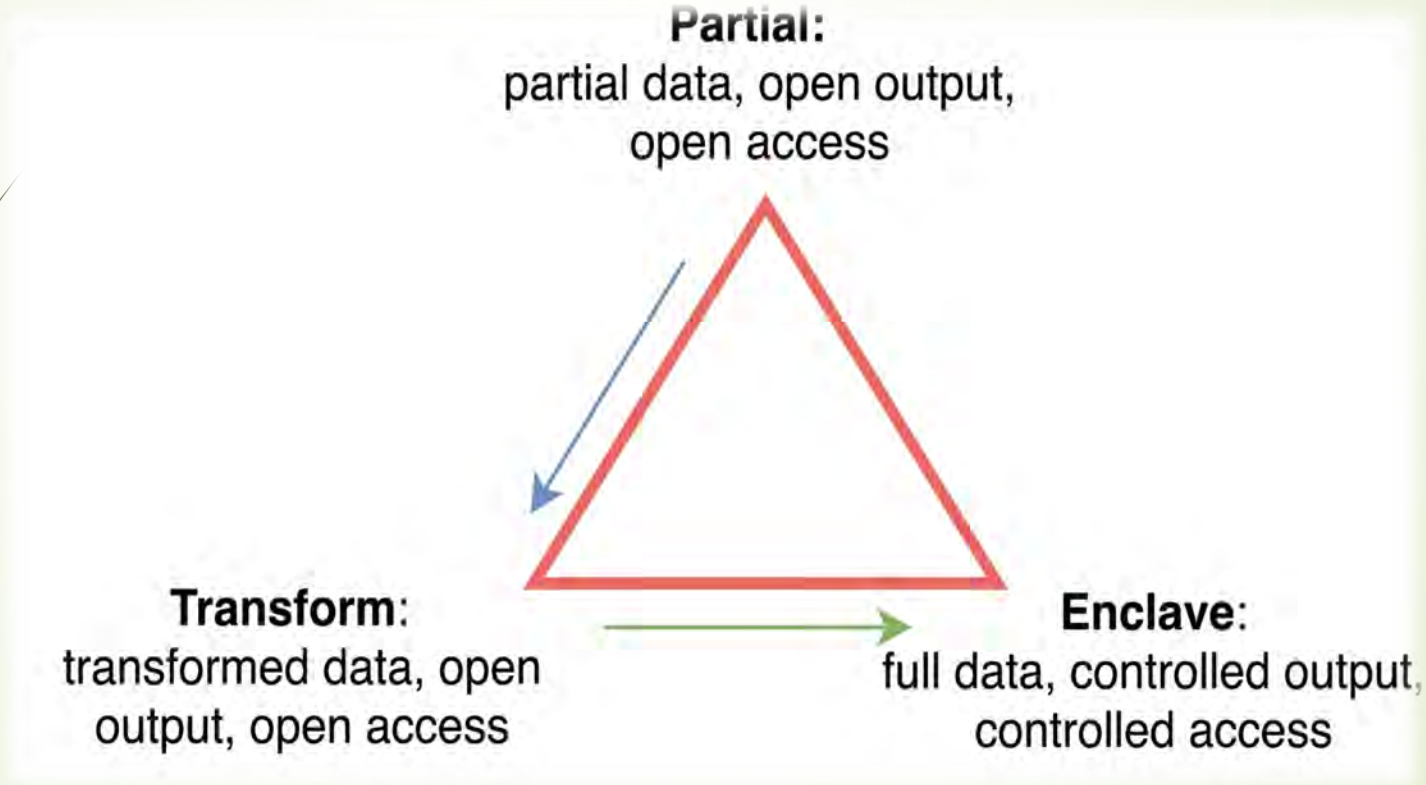
My study is on incarceration recidivism and employment in small towns

My data reveal locations of sensitive species

Open science is not open access; allows for "*open as possible, closed as necessary*"

Principle articulated in "Guidelines on FAIR Data Management in Horizon 2020", EU Horizon 2020 programme

Options for reuse of data "open as possible"



Socio-technical cyberinfrastructure

Synergies and tradeoffs exist between software components versus policy and process components in striking the right balance between safety for the data, ease of use, and efficiency.

Remote secure enclave consists of policies, human processes, and technologies that work hand-in-hand to enable controlled access and use of restricted data.

Plale, B., E. Dickson, et al., Safe Open Science for Restricted Data, to appear *Data Information Management*, DeGruyter Publisher 2019

Socio-technical cyberinfrastructure

Policies, human processes, and technologies that work hand-in-hand to enable controlled access and use of restricted data.

Socio-technical cyberinfrastructure must play role in reproducibility. What is minimal and sufficient role?

Peng, Z. and B. Plale, Queryable Citations for Reproducible Big Data Analysis, in preparation

Observation Takeaways

1. *Data valuation:* community norms in which data to keep and for how long
2. *FAIR:* PIDs are core to FAIR and to data reuse
3. *Open as possible:* cyberinfrastructure can and should contribute information for reproducibility of science that they support
4. *CI and reproducibility:* Not all science is built upon.

    E.g., Discard VM reproducibility bundle after 3 years

Beth Plale
plale@indiana.edu

Credit U.S. Library of Congress
Science grows by standing on the shoulders of giants