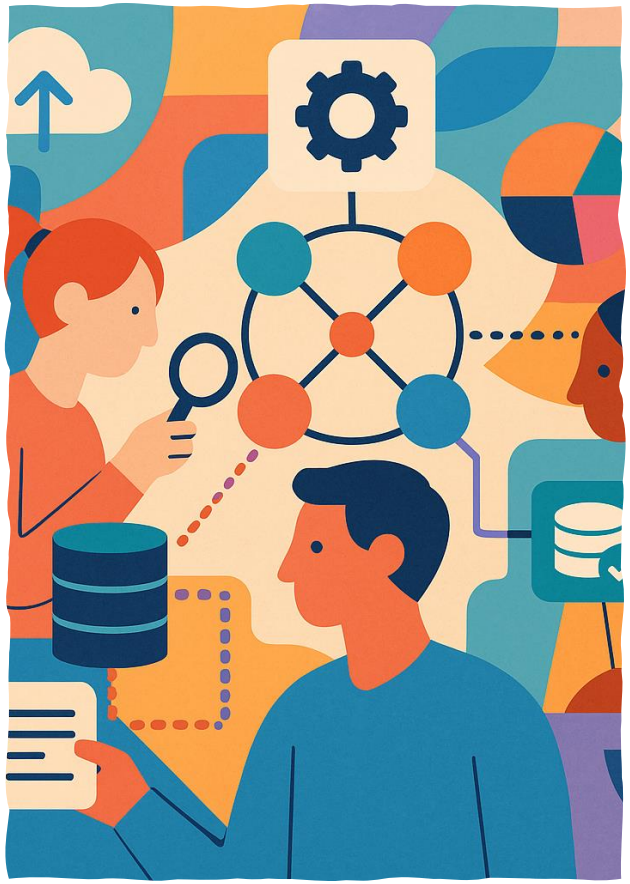
A Sample of Data Management Approaches From System Administrators and Researchers' Perspectives

Dr. Tabitha K. Samuel

Interim Director, National Institute for Computational Sciences,

University of Tennessee, Knoxville

Introduction



Research data management involves multiple roles:

- Data producers (instruments, codes)
- Data curators (e.g., librarians, analysts)
- Data operators (e.g., system administrators
- Data ingestors (e.g., researchers using data)

This presentation explores data management from:

- ❑ System administrator perspective
- ❑ Researcher perspective

System Administrators' Perspective

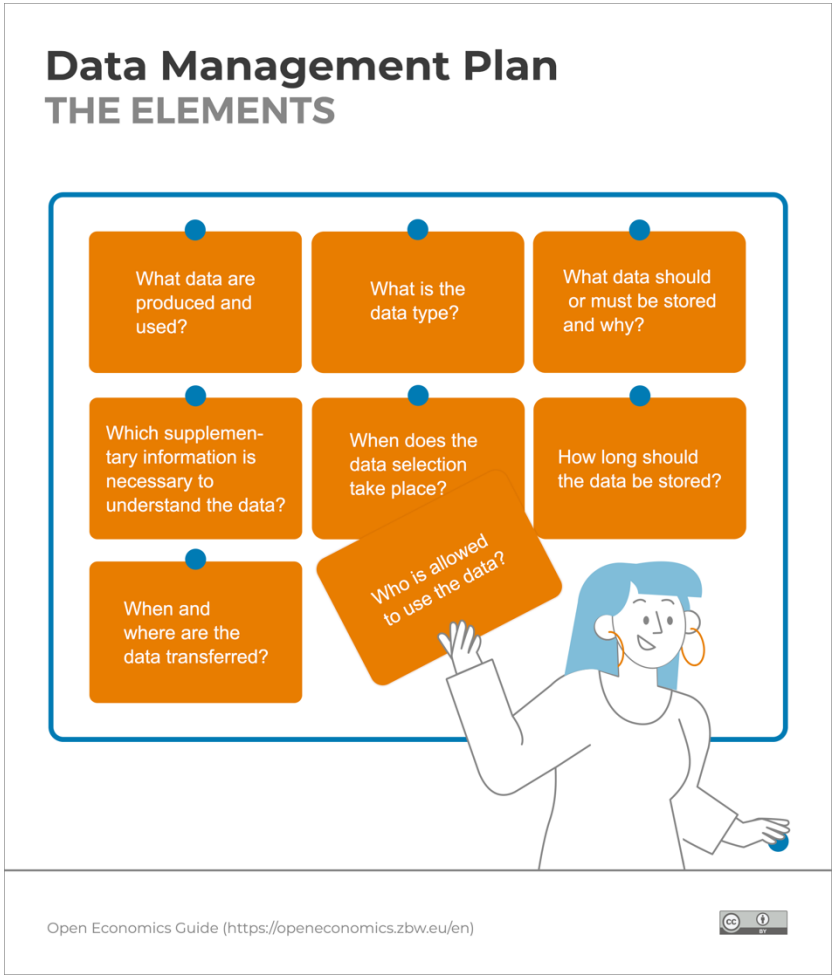
Quotas:
Research clusters have multiple storage areas:
- Home: config files, scripts (small quota)
- Scratch: temporary results (no quota, purged)
- Project: medium-term results (large quota)
- Archival: long-term storage (very large quota)

Tools to manage/report user data:
User portal
command-line custom tools
GUI one-stop shops such as OpenOnDemand

Administrators' Perspectives

- **Data purges**

- **Policies:**
  - Data management policies govern how data is ingested, stored, managed, archived, and deleted



Researchers' Perspective

The National Science Foundation recommends that open data have the following principles*.

* https://www.nsf.gov/digital/data

Redundancy and Ownership

Redundancy
- Archival storage for static data
- Backup systems for evolving data

Ownership
- Provenance and access control
- Determines storage and sharing policies

Data Sharing

Different rules for different data

Common Data Sharing tools
Command-line: rsync or scp
Cloud sharing platforms: OneDrive, Google Cloud, Box
GUI tools: WinSCP, FileZilla
High performance tools: such as Globus

Questions & Discussion