3

OBJECTIVES, ACTIVITIES, METHODS

The data curation objective

Data curation is concerned with all aspects of the management of data

in order to efficiently and reliably support the analysis of data, and enable reuse over time.

Areas of curatorial activities

Collection	Support the collection and acquisition of data
Organization	Employ an appropriate data model and use appropriate standards
Storage	Support reliable and effective storage
Preservation	Ensure that data will be understandable and useable in the future
Discoverability	Support the ability to search for and locate relevant data
Access	Support the ability to retrieve and distribute data
Workflow	Support the ability to systematize data workflows
Identification	Support the ability to identify, authenticate, and validate data
Integration	Support integration of data from different sources using different data models
Reformatting	Support reformatting for use by different tools or to match new format standards
Reproducibility	Support ability to reproduce results, ensuring scientific validity and reliability
Sharing	Support sharing data between researchers, teams, and institutions
Communication	Support representation, publishing, and visualizations that provide insight
Provenance	Support .identifying what inputs, processes, and calculations are responsible for data values
Modification	Support management of corrections and updates
Compliance	Ensure compliance to legal, regulatory, and local policy requirements
Security	Ensure that data is secure from tampering or inappropriate access and distribution

A Closer Look

Now let’s take a look at each of the areas. . .

Collection

Support the collection and acquisition of data

Includes support for, e.g., coordination of instrument calibration, protocols, procedures, collection area division, interview transcription, etc.

Of particular importance: recording information (as metadata) related to collection activity so that all relevant aspects of context are available later to support full understanding, authentication, and provenance.

Organization

Employ an appropriate data model and use appropriate standards

Determine an appropriate data model and schema

Use abstraction and indirection to manage data

Identify and use any relevant standards for both syntax and semantics

Of particular importance:

- Document schema attributes (including specifying datatypes and constraints).
- Document all changes to schemas.
- Maintain metadata for schema changes.

Storage

Support reliable and effective storage

Select storage strategies that proved the right mix of reliability, security and access

Preservation

Ensure that data will be understandable and useable in the future

Maintain a documented preservation strategy.

This includes not just bit sequence preservation and syntax documentation, but also the documentation of semantics for data elements and the generation and preservation of all metadata needed to ensure that the data is useable and understandable, and can be authenticated and audited for provenance.

Execute that strategy with discipline, documenting all actions taken.

Discoverability

Support the ability to search for and locate relevant data

Develop metadata to support searching for and finding relevant data in relevant formats.

Support searching that provide relevance ranking and recommends related datasets.

Access

Support the ability to retrieve and distribute data

Maintain systems, tools, and metadata that support the efficient and reliable retrieval and distribution of data.

Add metadata describing file formats

Where appropriate control access appropriately and maintain data on distribution and access.

Workflow

Support the ability to systematize work with data

The processing of data should be carried out a well-designed modular system of transformations.

The role of each module should be documented

The execution of a workflow should be documented as well.

To the greatest extent possible documentation should be generated automatically and should itself be both machine readable and executable.

Specifically: well-maintained scripts should be developed and used to document as well as execute data transformations.

Identification

Support the ability to identify, authenticate, and validate data

Identifier systems must be carefully designed.

Attention must be given to *what* (conceptually) is being identified and to the *method* of identification.

Related entities (such as the data abstractly and the same data represented in different formats) must be both precisely distinguished and precisely related.

Version control for format changes, corrections, etc. must be implemented.

Authentication (the data is in fact the data it claims to be) and validation (the schema constraints, syntax and semantics, are met) are both fundamental.

Integration

Support integration of data from different sources using different data models

Both variations in syntax and data element semantics must be accommodated if data from multiple sources is to be combined and related to solve real world problems.

Use schema alignment and cross-walking techniques to integrate data

Document integration strategies in detail so that any conflation, data loss, etc. is noted.

Reformatting

Support reformatting for use by different tools or to match new format standards

Data must frequently be reformatted in order to support new tools, new versions of existing tools, or to meet new format standards..

Reformatting must be documented and any changes in semantics or meaning must be identified.

Reproducibility

Support ability to reproduce results, ensuring scientific validity and reliability

Data curation for reproducibility includes documenting not only data collection and management, but also documenting processing and analysis.

Sharing

Support sharing data between researchers, teams, and institutions

There are many obstacles to data sharing, ranging from formats, to lack of documentation, to concerns about misuse or misunderstanding.

Data curation must address these, typically with policies, documentation, metadata, and interoperable systems.

Communication

Support representation, publishing, and visualizations that provide insight.

To be useful data must be presented in forms that provide insight (such as scientific visualizations) and integrated clearly and efficiently into the full life-cycle of scientific work, which includes scientific publishing. Related communication issues are relevant to other data curation activities: in entertainment, documentation, services, etc. Here data curation overlaps with interface design.

Provenance

Support identifying what inputs, calculations, and actions are responsible for data values

When one data set (or view) is derived from another, reliable use and understanding requires that the inputs, calculations, and actions responsible for data values can be identified.

Modification

Support management of corrections and updates

Data must be updated and corrected.

This must be supported and managed so that errors are not introduced but so that the changes overtime can be tracked and audited.

Compliance

Ensure compliance to legal, regulatory, and local policy requirements

The issues here range from intellectual property rights to regulations regarding the privacy of medical, financial, and personal information.

Security

Ensure that data is secure from tampering or inappropriate access and distribution

This will involve methods for controlling access and determining user identity and privileges, as well as data identity, authentication and validation.

Methods of curatorial action

Analysis: To determine needs, and develop relevant data models and *metadata*, and reformat, correct, or update data.

Documentation: To record essential information (typically via *metadata)*

System design and implementation: To support all data curatorial activities
To support the generation and use of data documentation and processing documentation

Policy: To specify objectives, procedures, practices, and formats.

Process: To ensure success and efficiency by managing the development of appropriate organizational units and roles, providing training, advocating for change, and managing curatorial activities.