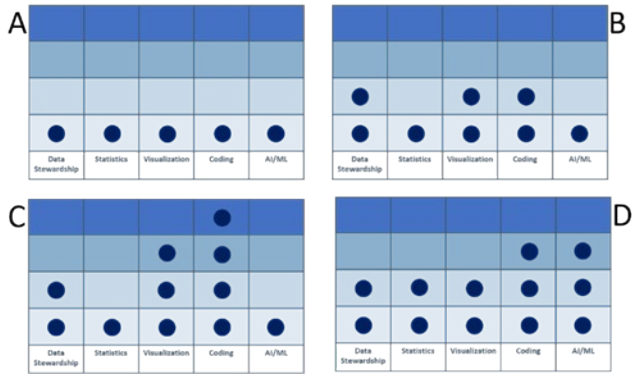
Example skill profiles

The figure below illustrates potential skills profiles that might be customized based on individual researchers' needs. A researcher who primarily runs experiments in a lab and is not responsible for data analysis may be certified at Essential level in all disciplines (A) and remain at this level until their role or career aspirations change. An individual who wishes to visualize and manage data and wants to learn to automate tedious tasks may desire to become certified at the Experienced level in Data Stewardship, Visualization and Coding (B). Someone in a role who performs complex coding and visualization tasks (such as a data analyst) as part of their daily work and who regularly helps others to improve in their coding skills may wish to become certified at higher levels in visualization and coding (C). A data scientist who regularly practices machine learning using programming tools like TensorFlow may become certified at higher levels in coding, statistics and machine learning (D).



Example of potential skills profiles based on a person's role.

Guidelines of CDS

The CDS program at Dow is based on a set of guidelines that are intended to ensure democratic, transparent data access, in general alignment with the FAIR data principles[1]. They inform the preparation and selection of training material. Note that these guidelines are written to be applicable even under extreme conditions, with a minimum of pre-existing professional IT support for the individual researcher. Moreover, they are framed for an environment where the research context (the "project") sharply defines both the subject area, the data environment, as well as the individuals requiring access to the data.

Data organization

1. *Store and organize data at the level of a project.*
   Keeping data in a (virtual) container that is separated from other data by project context facilitates findability, security, and allows to add project-specific meaning to the data. The intent of this guideline is to establish clear boundaries around a project's data that are easy to recognize and work with both for humans and machines (i.e., information systems) in the specific context of the project, while maintaining agility.
2. *Store raw data in a managed system.*

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18, see also https://www.go-fair.org/fair-principles/.

Storing a project's raw data in a secure, access-controlled, high-availability system that is under the control of the project's Data Steward[2].

3. *Organize data to reflect the observational hierarchy of the project.*
   The project's data workspace structure must be as intuitive as possible and should establish easily discernible patterns that facilitate composition and re-use, as well as place context around the data. This guideline naturally dovetails with the Tidy Data principles.
4. *Use Tidy Data tables*.
   The Tidy Data principles proposed by Wickham fundamentally enable agile, reproducible, and transparent research. The Tidy Data Format refers to:
   - Each row is an observation
   - Each column is a variable
   - Each table contains one type of observation
5. *Document the structure of the data.*
   The project's data must be as self-explanatory as possible to facilitate transparency and collaboration. For this to happen, the meaning of each observed quantity (e.g., what is being recorded in each table or column) should ideally be defined in such a way that the data becomes self-documenting.

Data access

1. *Make data accessible to both humans and machines.*
   Balance the benefits of automation (speed, reproducibility, and transparency) with the need to democratize data access for the project team: both humans and machines need to be able to access the data.
2. *Remove artificial barriers to data access.*
   Project team members should have a maximum amount of control over the data, as well as maximum effective bandwidth for interactions with their project's data. This ensures a constant, high level of engagement with the data and facilitates collaboration.
3. *Provide easy means to add and edit data.*
   It should be easy for project members to add new data to data sets and fix any errors that are found. Errors in the data need to be fixed as close to the origin of the data as possible and as soon as possible.
4. *Visualize your data as soon as possible, as much as possible.*
   Ensure frequent and thorough engagement of the team with the data. Engaging with data visually helps the project team and the data steward to illuminate various aspects of the data, drives cognitive processing of the data that in turn stimulates fresh thinking. It also provides a very effective means of error detection.

Data analysis

1. *Separate data storage and analyses.*
   The intent of this guideline is to ensure that data analysis methodology and procedures should be separated from the data they act on. This means that facts should go into tables and reasoning or methodology should go into code.
2. *Create transparent analysis routines*.

[2] The definition of "Data Steward" used here is somewhat modified from the common definition; "Data Steward" as used here refers to the researcher in charge of a project's data flow. It is not a compliance-centered role.

Transparency and reproducibility are ensured by creating well-documented or self-documenting pipelines that, when executed, perform a project's data analysis "at the click of a button".

3. *Divide analysis into minimal components with explicit documentation.*
   Dividing the analysis performed on a project's data into small, functional modules helps make work processes agile and easily reusable. All the procedures contributing to analysis (data harvesting, cleaning, transforming, modeling, prediction, etc.) should be as modular as one can make them.
4. *Implement version control for data, analysis & platform.*
   Reproducibility of a data work process is strongly increased by being able to trace and possibly revert to prior versions of the project's data or the procedures used to process and analyze that data. The intent of this guideline is to make ensure reproducibility by using change- or version control on project data and software.

Value preservation

1. *Package projects for long term storage.*
   This guideline's intent is to transcend single-use projects and to provide lasting benefit by archiving a readily re-deployable, self-explaining time capsule for the project's data useful in later research by other project teams.
2. *Transcend platforms and languages with shareable interfaces.*
   To avoid projects becoming information silos, and to minimize the effort of developing processes for new projects, all project data and procedures should be available across platforms and languages through the use of open standards and interfaces.