
Overview

This standard identifies the competencies you need to store, archive and curate data from life science-related experiments.

You will be required to demonstrate that you can design and implement a storage solution appropriate for your data, considering the type of data, access requirements and technical infrastructure. You must be able to process data into formats suitable for storage, and apply appropriate metadata or annotation to the data. Data storage solutions must comply to necessary data security and access restrictions in line with regulatory, legal or ethical requirements.

This activity is likely to be undertaken by individuals working in Life Science, Pharmaceutical, Chemical Biology, Agritech & Biotech industries. This could include job titles described as bioinformatics, computational biology, computational toxicology, Cheminformatics, Health informatics, Medical informatics, Agri-informatics for example.

Performance criteria

- You must be able to:*
- P1 process raw data from experiments into formats suitable for computational storage.
 - P2 evaluate the intended use of the data and select a storage solution suitable for this use.
 - P3 identify the IT resource, infrastructure and cost requirements of any chosen storage solution.
 - P4 collate suitable supporting information, metadata and annotation for storage with the data.
 - P5 design appropriate local and private storage structures.
 - P6 implement the chosen local and private storage solutions.
 - P7 ensure that data is backed up securely.
 - P8 secure access to the data in line with ethical and legal rules and guidance for the data type.
 - P9 secure access to the data in line with general organisational policies and commercial considerations.
 - P10 implement a data management plan appropriate to the data types you are working with.
 - P11 identify suitable public data repositories for data deposition.
 - P12 prepare, curate and submit data to appropriate public repositories.

Knowledge and understanding

- You need to know and understand:*
- K1 different types of biological and chemical data generated from life science experimental techniques.
 - K2 how to work with experimental colleagues to design experiments with suitable data collection and handling requirements.
 - K3 the limitations and complexities of life science data.
 - K4 file formats for raw data from experiments, including proprietary formats.
 - K5 conversion techniques to extract or process raw data.
 - K6 the data format requirements of downstream data analyses.
 - K7 the storage requirements of the data being collected.
 - K8 how to plan for future data generation and scale up.
 - K9 database design and management, including information security considerations.
 - K10 common database systems, including GDB and SQL.
 - K11 relevant big-data storage platforms.
 - K12 high performance computing platforms including Linux, Unix, local and remote HPC, and cloud computing.
 - K13 the cost of various data storage solutions – local and cloud-based options.
 - K14 data annotation standards and the required metadata for common life science data types.
 - K15 what public databases accept direct data submissions and where to access their submission procedures.
 - K16 ontologies and their use.
 - K17 the responsibilities of working in a production environment managing scientific data.
 - K18 current approaches for modelling and warehousing of life science data.
 - K19 the importance of data governance, curation, information architecture and ensuring interoperability.

Developed by	Cogent
Version Number	1
Date Approved	September 2018
Indicative Review Date	September 2022
Validity	Current
Status	Original
Originating Organisation	Cogent
Original URN	COGBIO-01
Relevant Occupations	Information and Communication Technology Professionals; Science; Science and mathematics Science; Science Professionals
Suite	Bioinformatics
Keywords	data governance, curation, information architecture, ensuring interoperability
