DATA CURATION
(CSEN 3144)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Which of the following are examples of quantitative data?
- (a) Hourly air temperatures in Fahrenheit recorded at Macalester's Ordway Field Station for the years 2010 - 2015
- (b) Photographs recording the starting stances of Olympic runners in the 2016 Olympic Games
- (c) Tally of the number of cars that passed through the intersection of Grand and Snelling Avenues last Friday between 7:00 a.m. and 7:00 p.m.
- (d) All of these.

(ii) A data curator typically identifies:
- (a) Required data sets and ensures they're collected
- (b) Data is cleansed and transformed as needed
- (c) Making the data sets and information about them, such as their metadata and lineage documentation, available to users
- (d) All of the above.

(iii) Point out the correct statement
- (a) Raw data is original source of data
- (b) Preprocessed data is original source of data
- (c) Raw data is the data obtained after processing steps
- (d) None of the mentioned

(iv) Among the following options choose which one of the following focuses on the discovery of unknown properties on the data
- (a) Big data
- (b) Data mining
- (c) Machine learning
- (d) Data wrangling.

(v) What are the three pillers of data curation?
- (a) Technology, Organization, resources
- (b) Organization, Expertise, Technology
- (c) Organization, Expertise, Resources
- (d) Resources, Expertise, Technology.

(vi) What is arXiv?
(a) Peer reviewed pre-prints for the field of physics
(b) Repository of electronic preprints of scientific papers in the field of mathematics, physics, quantitative biology and others
(c) A group of archives spanning many different scientific disciplines
(d) Preprint repository for the biological sciences hosted by Cold Spring Harbour Laboratory.

(vii) Which of these is the advantages for using a data management service (DMS)?
(a) Reduces data redundancy and inconsistency
(b) Manage data movement
(c) Promote data governance
(d) All of these.

(viii) Which of the following is the common goal of statistical modelling?
(a) Inference (b) Summarizing
(c) Subsetting (d) None of the above.

(ix) What is a structured representation of data?
(a) Database table (b) Functions
(c) Data prepration (d) Data frame.

(x) Qualitative data is also known as
(a) Numerical data (b) Categorical data
(c) Discrete data (d) Continuous data.

Fill in the blanks with the correct word

(xi) What does 'OAIS' stands for _____.

(xii) _____ is the process of moving data that is no longer actively used to a separate storage device for long-term retention

(xiii) _____ is a means of managing data that makes it more useful for users engaging in data discovery andanalysis.

(xiv) _____ is the process of moving data that is no longer actively used to a separate storage device for long-term retention

(xv) _____ is the process of importing large, assorted data files from multiple sources into a single, cloud-based storage medium.

Group – B

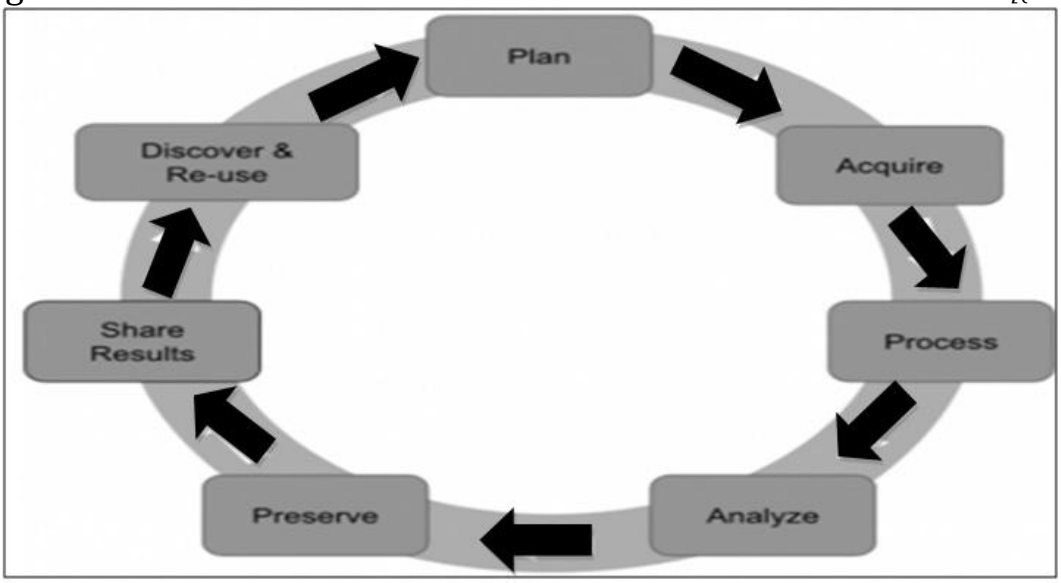
2. (a) What is Metadata Management? What is the Difference Between a Data Catalog and Metadata Management? *[(CO3)(Understand/IOCQ)]*
(b) What is the challenge in data sharing and what measure do you take before sharing the data in public. *[(CO2)(Remember/LOCQ)]*
(c) What are the benefits of managing research data? *[(CO2)(Remember/LOCQ)]*

(2 + 3) + (2 + 2) + 3 = 12

3. (a) Where does data curation fit into the data creation process? *[(CO1)(Define/LOCQ)]*

(b) What is the purpose of data curation? *[(CO1)(Define/LOCQ)]*

(c) 'Data management refers to the process of deciding and documenting how data will be collected, organized, stored and shared' - Justify this comment by the figure below: *[(CO1)(Define/LOCQ)]*



4 + 4 + 4 = 12

Group - C

4. (a) "The core impact of data curation is to enable more complete and high-quality data-driven models for knowledge organizations." – Justify this statement. *[(CO2)(Understand/LOCQ)]*

(b) What are the important phases of data life cycle model? *[(CO3)(Understand/IOCQ)]*

(c) What is the importance of data types and data formats? Give examples to support your answer. *[(CO4)(Understand/LOCQ)]*

(d) What is a discipline specific data format? Give an example. *[(CO4)(Understand/LOCQ)]*

3 + 4 + 3 + 2 = 12

5. (a) How your research datasets are described? *[(CO2)(Understand /LOCQ)]*

(b) How will the descriptive metadata be created or captured? *[(CO2) (Understand/LOCQ)]*

(c) With whom will you share your research data in the short term, before publication of any papers arising from their interpretation? Where will you store your data in the short term, after acquisition? *[(CO4)(Understand /IOCQ)]*

3 + 3 + 6 = 12

Group - D

6. (a) When will your research data be moved to a secure archive for long-term preservation and publication? *[(CO5)(Identify/HOCQ)]*

(b) How (i.e. by what physical or electronic method) will you transfer your research datasets to their long-term archive, under the curatorial care of a separate third-party, e.g. a data repository? *[(CO6)(Apply/HOCQ)]*

(c) Why is public access to your research data to be restricted (if indeed it is)? *[(CO5)(Identify /HOCQ)]*

4 + 6 + 2 = 12

7. (a) What's the Difference Between Restore and Retrieve? *[(CO1)(Define/LOCQ)]*

(b) What should be done to the data that was created from research which is no longer active? *[(CO4)(Understand/LOCQ)]*

(c) What is data ingestion? What are data ingestion challenges? How data scaling helps in dealing with such challenges? *[(CO1,CO2)(Understand/LOCQ)]*

3 + 3 + (2 + 2 + 2) = 12

Group - E

8. (a) The reference model identifies and describes the external entities constituting an OAIS's environment. Explain with a diagram what constitute the OAIS environment. *[(CO6)(Apply/HOCQ)]*

(b) When does a data requires a persistent identifier? *[(CO2)(Understand/LOCQ)]*

(c) What is the OAI in library science? *[(CO1)(Define/LOCQ)]*

6 + 3 + 3 = 12

9. (a) What is an example of a data type you keep for the very long term? *[(CO1(Define/LOCQ)]*

(b) Illustrate the application of 'arXiv' in Digital Curation, Digital Preservation and Data Management. *[(CO6)(Apply/HOCQ)]*

5 + 7 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	58.33	15.63	26.04

Course Outcome (CO):

After the completion of the course students will be able to

CSEN3144.1. Define and distinguish digital Curation, data Curation and related terminology.

CSEN3144.2. Understand the characteristics of various data types generated and used by a variety of fields, research communities, and government organizations.

CSEN3144.3. Understand the data curation lifecycle and identify the activities associated with each stage.

CSEN3144.4. Understand the importance of dataset identifiers, citation and data repository.

CSEN3144.5. Identify standards and technologies for managing and maintaining digital content.

CSEN3144.6. Apply theoretical understanding to practical issues in data Curation.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*