Purpose of this document

This document is a Q & A for past exam questions and and questions in the same style for FDM. The idea is to reduce 600 something slides to 20ish pages of relevant information for the exam.

There is one section per lecture. The titles follow the course outline on Learn. Past exam questions should be split so all questions in a section can be answered using the slides corresponding to the section.

Please add questions, answer them, correct typos, make this document better in any way. Should someone disagree on an answer I suggest discussing it in the facebook post where I shared this file. Answers should be as objective as possible and ideally just copy & paste from slides.

1 Introduction and motivation: data management issues in modern research

1. What can / should be published together with a paper? What are the advantages?

   Solution: The data, workflows and software used allow more effective scrutiny, reproducibility, validation and re-use.

2. Name some advantages of publically accessible research data

   Solution:

   - Increases the impact and visibility of research
   - Promotes innovation and potential new data uses
   - Leads to new collaborations between data users and creators
   - Maximises transparency and accountability
   - Enables scrutiny of research findings
   - Encourages improvement and validation of research methods
   - Reduces cost of duplicating data collection
   - Provides important resources for education and training

3. What is Big Data?

   Solution: Big Data means

   Volume Large files and many of them. Lots of bytes.

   Velocity Rapidly changing and growing data sets. Think twitter, streams from scientific instruments.

   Variety Complexity in data structures, unstructured data

4. How big is big in the web and science?

Solution: Facebook and Youtube store data in the order of 10 PB/year. Experiments like LHC and telescopes output processed data in the same range while raw data rates can be magnitudes higher. The output of computational sciences is limited by system memory (c. 100 GB) but ensemble runs can create multiple files. German climate modelling centre DKRZ also stores 10 PB/year.

2 Structured approaches to research data: overview, file formats, storage

1. Explain the three levels meaning is captured in digital data

Solution:

- 1. Encoding: How to record numbers and characters as groups of bits
- 2. File format: How to interpret the arrangement of the numbers or characters
- 3. Files/Records: How to collect bits into groups with meaning

2. What is encoding? Give examples.

Solution: Interpreting groups of bits; dpne at a character/number level (ie 0 or 1)
eg encoding formats for fixed bit lengths - i.e 8 bits ⇒ 1 byte ⇒ 1 ASCII character.
Computers work with bytes as a base unit ⇒ 4-byte integer rather than 32-bit integer

3. What is a file? What is digital forensics and how does it work?

Solution:

- File: Basic unit of data organisation, i.e pointer to a sequence of bits and a (small) collection of descriptive information (metadata) about that sequence
- Digital Forensics: Use of tools to recover digital data from files or disk images (e.g disaster recovery/ law enforement or investigation)
  - work on the principle that deleting files doesnt actually delete data, deleting just removes link from inode table... bit sequence still on disk until overwritten

4. What is a file system? Discuss examples.

Solution:

- File System: framework for creating, manipulating, deleting & keeping track of files
  - overlays the actual data stored on disk / tape / flash drive
  - logical view of the data and their organisation

- Examples:
  - Desktop: Windows (FAT, NTFS); MacOS (HFS); Linux (ext2, ext3, ext4, xfs, btrfs)
  - Network/Distributed: Linux/Unix (NFS); Windows (SMB); MacOS(AFP)
  - HPC: Linux (ext3, ext4 + NFS); HDFS (hadoop?)
- Modern computer systems support a wide range of file systems as well as their native ones

5. Why are Fortran generated binary files a bad idea?

Solution: FORTRAN i/o is record-based

- binary data in array is written as a record, bracketed by its size in bytes
- implementation dependent
- FORTRAN-created binary files are distinctly unportable ⇒ avoid using them

6. What forensic tools do you know?

Solution:

- Emacs
- Open source forensic tools: Sleuth kit; Foremost

3 Structured approaches to research data: relational databases

1. It is a common practice in industry to use a Database Management System (DBMS) to organise and manage data. *Exam 8.12.2014 Q 4.a.-b.*
   (a) List 5 things that a DBMS provides.
   (b) Relational databases are widely used to manage data.
      i. What is a relational database?
      ii. Briefly explain the 4 ACID properties of a transaction.

Solution: A DBMS provides

- Mechanisms to create data structure (e.g., Tables) and content
- A means of querying and modifying content
- Ways to optimise performance
- A tool to backup or archive
- Ways to allow applications to access data

A relational database is a collection of tables (i.e., Relations) with associated relationships.

Atomicity All commands of a transaction is performed or none of them

Consistency A transaction takes a database from one consistent state to another consistent state.

Isolation When transactions are executed in parallel, the effects of one transaction must not visible to another transaction until the transaction is committed.

Durability After a transaction is committed, the changes made by that transaction is permanent. This is important, for example, if a system failure occurs.

2. What is a database?

Solution: A means of storing and accessing data efficiently. Contains a database management system (DBMS).

3. Define what is
   i. a primary key.
   ii. an index.

Solution:

i. A primary key is a field that is guaranteed to provide a unique value in each table.

ii. An index is a method of accessing entries in a table.

4. What constraints does a relational database have? Briefly describe each.

Solution:

- Domain Constraints : value for a field must be picked from a particular set (or range) of pre-defined values
- Uniqueness Constraints : If a field is a key, each record must hold a unique value for that field
- Null Constraint : whether or not a field can hold `NULL` as its value
- Referential integrity : foreign keys are references to external tables; they must correspond to valid primary keys in another table. If a record is deleted in a table, any other which references that records primary key as a foreign key must also be deleted.
- Semantic constraints: e.g salary of an employee must be less than that of his boss
- Dynamic constraints: e.g yearly salary can only increase

5. What methods exist of increasing the performance of a database?

Solution:

- Use meaningful indexes, e.g primary key attached to one record.
- Write data as columns to be used by these indexes. Columns make reductions (e.g SUM, MAX) faster

6. What are ER-diagrams and how can they be useful?

Solution:

- Entity-Relationship: Maps which show the relationships between objects (see slides for picture)
  - Binary relationship: Manager `manages` Employee
  - Tertiary Salesman `sells` product `to` customer
- Useful to produce (and visualise) representation fo real-world situation
  - Identifies important entities (tables)
- Simplification of model
  - E.g., Remove many-to-many relationships

7. Querying Relational Databases

Solution: Use Structured Query Language (SQL) ⇒ Used by all major DBMS's; standard versions exist

- SQL query of a table ⇒ Results in another table
- example query:
  `SELECT` stuff
  `FROM` table
  `WHERE` condition 1 `AND` condition 2 `OR` condition 3
  `ORDER BY` field 1 `DESC`
- Join tables together using `JOIN`, based on shared fields
  `SELECT` stuff
  `FROM` table 1 `JOIN` table 2
  `WHERE` table1.field1 = table2.field2 `AND` table1.field3 != table2.field4
- Can use pattern matching (e.g wildcards, `%`) and aggregation (e.g `SUM, MIN`), and ordering (e.g `ORDER BY` names [alphabetical]; `ORDER BY salary` [from low to high]).

4 Structured approaches to research data: NoSQL databases

1. It is a common practice in industry to use a Database Management System (DBMS) to organise and manage data. *Exam 8.12.2014 Q 4.c.-d.*

(a) Explain the 4 data models in NoSQL databases. Mention at least one DBMS for each data model.

(b) List 3 differences between relational and NoSQL databases.

Solution: Data models:

Key-Value Data is stored as key-value pairs, i.e. hashtable that is persistent in disks. Key is typically number or string. Value can be anything. Amazon DynamoDB is an example.

Document Collections of documents that have a unique ID are stored. Schema-less, gives flexibility. Can be used similar to Key-Value stores but documents can be searched. JSON documents with nested key-value pairs are common. MongoDB is an example.

Column Columns are treated individually: Allows processing as array, Column values are stored contiguously, Faster processing of aggregated functions. Rows (tables) could be constructed from column values. Examples: Googles BigTable, Amazon SimpleDB, Apache Cassandra.

Graph Nodes represent entities with properties: E.g., ID, Name, Age, etc.. Edges represent relationship e.g., ID, Knows, Member. Each node and edge has a unique ID and knows adjacent nodes. Example: Neo4J.

Differences

- NoSQL has no joins
- NoSQL has no schemas
- NoSQL has no transactions

2. What is NoSQL? Discuss advantages and limitations of its implementation.

Solution: Not Only SQL ⇒ non-relational databases

- Other types of database, e.g XML/JSON based document based databses; Graph databases
- Designed for distributed storage with high horizontal scalability
  - for large structured, semi-structured or unstructured data
  - applications can store objects without using things like Object Relational Mapping (ORM)
- Does not impose schema on data ⇒ More flexibility
- Cannot perform joins across datasets
- Cannot perform transactions

3. What is MongoDB? Discuss advantages and limitations of its implementation.

Solution: Open Source JSON-like document based DBMS

- documents stored in binary JSON (BSON) format

- supports C, C++, C#, Java, Python, PHP, Perl, Ruby,...
- High performance ⇒ fast
- various data replication strategies possible

4. How can data in MongoDB be accessed?

Solution:

- Documents contain collections of key-value pairs
- If no primary key provided, MongoDB creates one automatically
- Accessed through MongoDB Shell - command line tool; or else interactive JavaScript application
- Allows 4 fundamental operations, CRUD:

  Create Insert new JSON-like documents to a collection

  Read Retrieve existing documents in a collection

  Update Change existing documents in a collection

  Delete Remove documents from a collection

5. How can you model data in MongoDB? Compare normalised and denormalised data models.

Solution:

- Normalised data model
  - No duplication of data
  - But need several requests for update..
- De-normalised data model
  - Duplication of data ⇒ application may need to define ways to deal with duplicate data
  - All relevant info in one place ⇒ one update step

5 Structured approaches to research data: XML and JSON

1. (a) List some components/syntax of XML and briefly describe each.

Solution:

Declaration: `<?xml version="1.0" encoding="UTF-8"?>`. Optional parameter useful for a parser.

Tags: Text between <and >, used for structuring documents.

Elements: Blocks of code between a start tag and an end tag. Self contained with <self/>. One root element per XML doc.

Attributes: Name-value pairs that provide information about an element. Attribute names are unique within the same element.

Comments: Called using <!– and –!>. Ignored by parsers.

(b) What is the purpose of the following when talking about XML documents? Briefly describe each.

i. Schemas

ii. Parsers

iii. XPath & XQuery

iv. Data Persistence

Solution:

i. Used to validate XML documents, i.e., give them structure. Some industries have standardised schemas.

ii. Document Object Model (DOM) and Simple API for XML (SAX).

iii. **XPath:** Allows navigating XML documents.
XQuery: Allows querying XML documents.

iv. Stores application configuration and state.

2. (a) Repeat question 2 (a) for JSON.

Solution: Much simpler than XML, using blocks to denote data elements.

(b) How can JSON Schemas be used for Data Validation?

Solution: Similar to XML Schema, a JSON schema describes valid content for application/domain specific JSON documents

6 Metadata

1. It is widely accepted that there is value in re-using research data. *Exam 8.12.2014 Q 2.a.1-2*

(a) How can metadata help a researcher re-use data?

(b) What properties does good metadata have?

Solution: Metadata makes data more re-usable because it makes it more understandable. It can be reusable for the same purpose (e.g. to aid validation of the results) and potentially others. It facilitates finding related data.

Metadata is good if it allows your data to be found and understood by all those who might want to make use of it. It should be:

- Accurate

- Precise
- Conforming to standards
  - Semantic: Meaning of Terms
  - Which metadata are mandatory
  - Formatting / Syntax
- Accessible
  - Online, addressable (can be linked to), harvestable

2. What is system Metadata? Content Metadata? Examples.

Solution:

- System/Structural Metadata:
  File ownership, modification date, how its packaged, etc.
- Content/Descriptive Metadata:
  - What/where/when/who the data relates to
  - how/why was the data collected
  - who collected the data; where was it collected; when was it collected

3. Where is metadata stored?

Solution: Can be emebedded alongside data, or in separate metadata files/indexes/catalogues

4. What are semantics? What is semantic annotation?

Solution:

- Semantics are the meaning of the data
- e.g specific meaning of a "Date" column ⇒ could refer to date data relates to, was created, was stored.. could be time/day/month/year...
  ⇒ Need to know more about a piece of data to understand its relevance
- Concepts used in the data may need clarification (e.g if "rain" includes "sleet" etc

5. What is an ontology?

Solution: A controlled vocabulary

- Used to describe semantics

- definitions for terms in the context of data/format
- used in metadata and data

6. Who should create metadata? What metadata do you want when you send or receive data?

Solution: Data author should create metadata since they know most about what the data represents. Anyone who understands it fully may contribute also.
Any information which makes the data more understandable/useful should be included when writing metadata.. what do you want, e.g:

- When sending
  - why were data created
  - limitations of data
  - what does data mean (content metadata)
  - how to cite/ licensing
- When receiving/resuing data
  - what are the data gaps
  - what methods used to generate data
  - fees/licencing? how to cite
  - scale of data
  - what do values mean (content metadata)
  - what software is needed to unload/read the data
  - can I redistribute this data

7. How do you choose a metadata standard? Name and describe a few.

Solution: Your choice of metadata standard may depend on your field of practice and your motivation for using metadata. Initiatives such as the RDA's **Metadata Standards Directory Working Group** are working to help people find the right metadata standards for them to use.

Dublin Core Element Set: For web resources and publications

FGDC* Content Standard for Digital Geospatial Data: Geospacial data with profiles and extensions to biological data

ISO 19115/19139: Geospatial data and services

Ecological Metadata Language: Ecological Data

Darwin Core: Emphasis on museum specimens

Geography Markup Language: Geographic features (roads, highways, bridges)

7 Data management & movement in HPC: iRODS and gridFTP

1. What is iRODS? Why is such a system useful?

Solution: iRODS is a data management system widely used (incl. EPCC)

- Open Source, free and actively developed and supported
- **I**ntegrated **R**ule-**O**riented **D**ata **S**ystem
- It is developed and supported by the iRODS consortium

2. What are the main features of iRODS?

Solution:

- Supports large numbers of users and user groups in a single data grid
- Supports heterogeneous data storage resources (Unix File Systems, Amazon S3 Buckets,..). More being developed
- Files exposed to user in a single unified namespace
- Handles peta bytes of data
- high performance network data transfer protocol, parallel I/O, similar to gridFTP
- metadata catalogue iCAT manages access control and mappings between logical and physics name spaces
- security iRODS username / passwords, Pluggable Auth Modules (PAM) for LDAP, Grid Security Infrastructure (X.509), Kerberos, Shibboleth
- rule engine allows automation of operations (valdiating, backing up, logging,..), enforces data mangement policies (privacy, retention) and enables rule based workflows
- Data Grid Federation: independent data grids can be federated with one another. Allows remote access to grids operated by separate workgroups
- Clients: CLI, Web and Desktop GUI clients, APIs in many languages

3. What can iRODS do?

Solution:

- Data Centre mangers: Simplifies data grid management
- Users: Simplifies data discovery, validation and processing
- Data Preservation - Digital Archives
- Data Maintenance

- Data Sharing and Access
- Policy enforcement
- Data Protection and Security
- Data Curation - Digital Libraries
- Automated Data Processing
- Distributed Data Management

4. What is GridFTP and what it is used for?

Solution:

- Extension of the standard File Transfer Protocol (FTP) defined by Open Grid Forum
- Widely used by many HPC centres (incl. EPCC)
- standard protocol for data transfers
- reliable, high performance
- can transfer large files over Wide Area Networks (terabytes range)
- Free & free clients

5. What are the main features of GridFTP?

Solution:

- Security with Grid Security Infrastructue (GSI using X.509)
- Third party transfers: Transfer between remote servers initiated by local client
- Parallel TCP streams: Higher use of bandwith by allowing multiple streams
- Striped data transfes: Used for clusters with parallel file systems. Each node reads a sections and sends it. Works together with parallel streams.
- Partial file transfers: Allows transfers of sections by specifying offset and block length, useful when only a small section is needed
- Fault tolerance and resuming transfers
- Clients: CLI (globus-url-copy, UberFTP) and Web (Globus File Transfer)

6. Explain Globus File Transfer

Solution:

- Globus File Transfer is one of the Globus services
- Can be accessed at globus.org
- Underlying technology is GridFTP
- Software as a Service hosted by Amazon Web Services (AWS)
- Data is transferred between Globus endpoints which are GridFTP Servers
- Personal computer can be turned into an endpoint using Globus Connect Personal
- Easy to use fire and forget mechanism for data transfer via web browser and command line. Can also be embedded via APIs

8 Elements of data management planning

1. Sketch a typical digital data lifecycle model. *Exam 8.12.2014 Q 1.a*
   (a) Why is such a model useful when dealing with digital research data?
   (b) Which parts of the lifecycle should a data management plan address?

Solution: A DMP is useful because

- It saves time and reorganisation later
- It increases research efficiency
- Many funding agencies require it
- Makes data preservation easier
- Prevents duplication of effort
- Can lead unanticipated discoveries
- Increases visibility of research
- Makes research and data more relevant

A DMP should adress all parts of the data lifecycle. An example of stages in a full life cycle would be: Plan, Create, Assure, Describe, Preserve, Discover, Combine, Process.

2. Describe the stages of a data lifecycle

Solution:

(a) Plan: **Data Management Plan (DMP)**
   - See DMP below

(b) Create
   - Observe, measure, generate by simulation

- ⇒ Lab notebook, organise it from the start

(c) Assure

- validate, calibrate
- Check correctness of methods/code
- Record calibration methods used (e.g in case raw data needed to be corrected to account for instrument bias)

(d) Describe

- Use meaningful variable names, record units used (metres/centimetres etc)
- record necessary info to be able to interpret the data
- conform to metadata standards

(e) Preserve

- Data now part of the scientific record ⇒ store them appropriately
- Consider backup/replication; accessibility (archiving vs darkiving); keeping data and its metadata together
- conform to metadata standards

(f) Discover

- Data must be findable ⇒ how do you make yours disoverable
- How do you discover other data that may save effort/duplication
- Consider description and accessibility

(g) Combine

- combine/integrate/merge data to create new insights
- requires good metadata to understand other data; require good tools
- be wary of licensing conditions ⇒ can you use this data in this way?

(h) Process:

- Use software to generate new data from old data
- Analysis of digital sensor data; simulation input; re-analysis of integrated third-party data

3. Describe the content of a DMP

Solution:

- what will I do
  - what data will I create?
  - how will I descibe them?
  - How will I store them?
  - Will they be published? ⇒ why/why not?

  - how can others find them?
- what is data management plan (DMP)?
  - Formal document to answer these questions
  - outlines plan for during and after research
  - ensures data is safe now and in future ⇒ Longevity of data
- Why prepare one?
  - save time later ⇒ already made design decisions; simplifies data preservation; prevents duplication
  - increase research efficiency ⇒ data will be available and understandable for future use. Increases research visibility. Makes research & data more relevant
  - Required by many funding bodies

The Standard components of a general DMP are:

- Information about data & data formats
- Metadata content and format
- Policies for access, sharing and re-use
- Long-term storage and data management
- Budget

9 Structured approaches to research data: RDF and semantic data

1. *Exam 8.12.2014 Q 2.b.*
   (a) Describe the main features of the Resource Description Framework (RDF) data model.
   (b) Pick out two aspects of semantic web technologies and explain how they could be used in the context of metadata and persistent identifiers.

Solution:

(a)
- RDF is a data model for representing semantic data
- It relies on giving every resource (every thing) a unique label (a URI) and describing relationships between resources with triples: Subject Predicate Object
- RDF can be serialised in multiple ways

(b) **Metadata**
- Allow data linkage in order to make related data more findable.
- Gives structure to data in the web.
- Allow data to be readable and understandable by humans.
- Larger quantities of metadata allow broader searches to link to one resource.

PID

- Introduce URIs in order to make resources more easily identifiable.
- Semantic web requires concise and unabmiguous persistent identifiers.
- Make data unique and decrease the chances of mistakes being made when finding and identifying data.

Web Ontology Language (OWL)

- Family of knowledge representation languages for authoring ontologies.
- Defines standards for metadata ⇒ consistency in terminology across datasets

2. How can Resource Description Framework (RDF) be serialized?

Solution:

Turtle: Subjects and objects are resources, predicates are parts of an ontology. e.g., Scotland¿ Capital¿ Edinburgh. Uses prefixes; rdf, rdfs, owl,xsd, dc, foaf, to categorise database elements.

N-Triples: Subset of Turtle, no linebreaks in a triple, no compact URIs.

N-Quads: Superset of N-Triples, subset of Turtle.

JSON-LD: A JSON file that contains RDF data.

N3:

RDF/XML: An XML document that contains RDF data.

3. How is RDF stored?

Solution: RDF is stored in files that use serialisation such as RDF/XML or Turtle.

4. What is a RDF vocabulary?

Solution: An RDF vocabulary is a collection of IRIs intended for use in RDF graphs. For example, the IRIs documented in [RDF11- SCHEMA] are the RDF Schema vocabulary. RDF Schema can itself be used to define and document additional RDF vocabularies.

5. How can RDF be accessed and queried?

Solution: RDF can be accessed and queried by something like one of the following

- Downloading bulk RDF
- Web query (curl)
- SPARQL

10 Persistent identifier systems

1. It is widely accepted that there is value in re-using research data. *Exam 8.12.2014 Q 2.a.3-4*
   (a) How do persistent identifiers help researchers make their data re-usable?
   (b) What makes a good persistent identifier?

   Solution:

   Objects needs to be (globally) addressable so that they can be reusable. Once data is addressable, it can be found more easily, can be cited and can be linked together.

   A good persistent identifier is not based on any changeable attributes of the entity, opaque (preferably a dumb number), unique. Human-readability, paste-ability and fit to common systems (e.g. URI) are nice to have.

2. What are PIDs? What are the properties of PIDs?

   Solution: PIDs are unique identifiers for data sources.

   They are unique, persistent over time, and they establish a redirection layer.

3. Give advantages and disadvantages of PIDs

   Solution:

   Advantage: Persistent identity vie indirection, persistent, inheritable across a broad set of entities. Searchable/ Querable.

   Disadvantages: Extra layer of effort to create and maintain, will the cost of a PID benefit you in any way? Persistant effort required to keep it updated.

4. Give example of PID systems

   Solution: There are several different PID systems/infrastructures

   - Handle System
   - Digital Object Identifier (DOI)
   - Archival Resource Key (ARK)
   - Persistent URL (PURL)
   - Life Science Identifier (LSID)
   - Uniform Resource Name (URN)

5. How do PIDs establish a redirection layer for data?

Solution: PID points to a URL, which points to a digital object. If the digital object or its attributes are changed, often the URL is also changed. The PID can be made to point to the new URL.

6. Why do we need to use PIDs for data?

Solution: Research data needs to be globally addressable using PIDs because

- It can be found easily by other researchers (meta search)
- It can be cited in research publications through a common reference label
- It can be linked with other similar data sets.

7. What makes a good PID?

Solution:

- Not based on mutable attributes (host/ org name etc).
- Opaque (random string/ numbers)
- Unique
- Human readable
- Copy-pasteable

11 Publication and citation of research data

1. Meaning of publish

Solution:

- Make it available for others to use
  - Either openly available or with access controls
  - stored with third party for long term preservation
- Make it citable
- possibly peer reviewed
- Publish alongside scientific paper; into long term repository (e.g EUDAT)
- Increasingly mandatory with funding

2. What is citation; where does it fit in Data Life Cycle; benefits

Solution:

Citation Providing a reference to data in the same way as researchers provide a reference to printed resources. **Author** is the researcher/individual who generates the digital data. Should have a persistent identifier (which directs uniquely to the cited data)

Life Cycle In the Describe section

3. List some properties of good Data Citation practices

Solution:

- Unique Identification
- Persistance
- Access to underlying data
- Specificity & verifiability
- Interoperability & Flexibility

4. What colaboration is required between Data Authors and Publishers (journals)

Solution:

- Citation standards (format, metadata, location, weight)
- Citation of prior data sets
- Preservation timeline and policies

5. Benefits of publication/citation?

Solution:

- Short Term
  - Discovery of relationships between data and publications ⇒ validate and build on previous work
  - Credits author ⇒ encourages proper publishing/citation in future
  - Link to publication provides information about related methodology and allows data to be put into context
- Long Term
  - measure impact of dataset and data author over career, same as papers currently measure researchers
  - ensure longevity of data

- more searchable data ⇒ Less *stealable* by another researhcer
- save duplication ⇒ Data needed in future may already exist
- Transparency in scientific research ⇒ increased rate of research

12 Archiving and preservation: maintaining the scientific record: OAIS model & repositories

1. (a) List some preservation media and some of their advantages and disadvantages?

Solution:

Stone: Pros: Durable, Cons: Slow Read/Write, low capacity

Wax: Pros: Durable, Cons: low capacity, slow read/write, fragile

Paper: Pros: Durable, Available. Cons: low capacity, slow read/write, damaged by elements

Magnetic Tape: Pros: Fast read/write, high capacity. Cons: Prone to mechanical failure

Optical: Pros: Fast read/writen, durable, good for long-lasting storage. cons: Capacity not good for large volumes of data, not good in fire.

Magnetic disc: Pros: Very fast read/write, high capacity. cons: not good around magnets or fire, break when dropped.

Solid state: Fast read/write, high capacity, durable. Cons: Magnets, fire, QM?

DNA: Pros: very high capacity, very durable. Cons: slow read/write.

2. *Exam 8.12.2014 Q 1.b.*
   (a) Explain what the OAIS model is, and sketch its major components.
   (b) For which stages of the data lifecycle model would an OAIS-compliant system be relevant?
   (c) Explain the differences between an Archive Information Package and a Dissemination Information Package.

Solution: Open Archival Information System reference model: A standard model of a digital archive

OAIS can be useful in the preservation stage of a data lifecycle.

AIP: Archive Information Package: how the archive stores data. AIPs are containers which bind together data objects, RI, other metadata including archive management details (timestamps, rules for copying, etc). AIPs are internal to an archive

DIP: Dissemination Information Package: how data objects are presented to consumers from the Designated Community (or elsewhere). These are logically distinct from AIPs but are typically derived from them

3. What are difficulties in planning long term preservation?

Solution: Data must be accessible, readable, and easy to descipher once you have access, not just by yourself but by other researchers. It must also be stored in a place where it is guaranteed to still be after a long time. Moreover, it is preferable to store that data that you know will be useful to you.

4. What concepts does OAIS use to plan long-term preservation?

Solution: Using the model and ideas from OAIS helps a lot in data management planning on any scale since:

- It provides a community of reserachers to review and scrutinise your data
- It requires metadata to, and representative information (RI) in order for the data to be understood.
- In creates a Submission Information Package (SIP) to the data creator containint the object and all the necessary representation information
- In creates an Archive Information Package (AIP) that explains how the data is stored, that is for internal use to the archive. AIPs are containers that bind together the data objects, RI, and other metadata.
- A Dissemination Information Package (DIP) is created for anyone that wishes to use the data. It is how data objects are presented to the consumer. They are derived from AIPs but are logically distinct from them.

5. What are advantages and disadvantages of distributed and unmanaged archiving?

Solution:

- **WWW**

  Pros: robust, resilient, protocol-centric, easy to publish

  Cons: how do we find it? is it quality controlled? how stable is the location of your data?

- **P2P**

  Pros: highly distributed, shared by peers

  Cons: searching (some trackers are blocked by providers), quality control, versioning?

6. Explain the role of each phase of the OAIS model (Ingest, Data mgmt, ...)

Solution: An OAIS model is essentially split into 6 autonomous cogs, where the internals of each cog in the wheel is invisible to the other parts, and the different parts communicate with each other by passing specific messages like SIP, AIP and DIP.

- Ingest

  - Gather the Data objects (DO) and Representation Information (RI). This can be done through the SIP submitted by the Data owner, and separate information gathered by the Ingest team
  - Complete the metadata info
  - Wrap the DO and RI into a AIP
- Data Management
  - Recvs Descriptive info about the data
  - Categorize and track all DOs (all copies) in the archive
  - Provide queryable/ searchable info on the documents. The queries will be run on the meta info for each DO.
- Archival Storage
  - AIP is stored here
  - Will implement policy based storage of AIP - replication, timeline
  - May alter storage methods in the future without informing other parts of the system
- Access
  - User submits the query to interact with the Data mgmt sys
  - Runs the searches, and retrieves the appropriate AIP from the storage
  - May repackage the AIP into a DIP which is more consumer friendly format
- Preservation Planning
  - Policies for Data management (replication, timeline, etc)
  - Usually human driven
- Administration
  - Overall monitoring of the flow of the data from Ingest to Access
  - Automated usually

7. List two examples of Data management systems built on the OAIS model

Solution:

- DSpace - Datashare (Univ. of Edinburgh) is based on this
- Fedora

13 Legal aspects: ownership, copyright, licensing, certification

1. What issues are faced when dealing with the differences between scientific and legal motivations?

Solution:

- Science is about community, law is about ownership.
- Science is about disclosing data, law limits the availability of data.
- Science promotes originality and scepticism, law works to provide regulations and conventions.

2. (a) What parts of data are protected by copyright? What parts aren't?
   (b) How long does copyright last?
   (c) What does copyright grant you?

Solution:

(a) Scientific, literary, artistic works, and their derivatives are protected. Ideas, raw facts, and mathematics are not protected. It is a key measure of originality and is attributed automatically to the creator upon creation of data.

(b) In the EU and in the US: 70 years after the death of the author (min 50 years in other countries)

(c) economic and moral rights.. i.e you have the right to

- make copies of it
- make it available to the public (eg. via publication or upload)
- make derivative works
- assert paternity of it
- disclose it
- have it respected

3. What are Database rights? How are they usedin science?

Solution: Database right prtoects

- work/data/materials arranged systematically and individually accessibly by electronic or other means
- qualitatively and/or quantitatively substantial investment in either the obtaining, verification or presentation of the contents

Scientific measurements are not covered by intellectual property rights ⇒ collection and organisation may be protected
Database rights prevent

- extraction/reuse of all/substancial parts of a database
- repeated & systematic extraction/reuse of insubstancial contents of a database if against the usual (licenced) exploitation of that database; or for use which are against the wishes of the data author

4. What are licences? How are they used?

Solution: To *license* is to grant someone official permission.
A *Licence* is an official document which grants permission. Licences allow you to use others' work subject to conditions. Can license Copyright and Database rights. If no licence available, data are either public, or you don't have permission to use

5. (a) What aspects exist for Creative Commons (cc) licenses?
   (b) Which of these aspects are non-open?
   (c) Name some users of Creative Commons licenses?

Solution:

(a) Attribution (BY), No Derivatives (ND), Non Commercial (NC), Sharealike (SA), Public Domain.

(b) No derivatives, non commercial.

(c) Flickr, Google, OpenCourseWare, Wikipedia, Radiohead, White House

6. List some other open public licenses.

Solution: Open Database Licence (ODbL); Open Data Commons Attribution Licence (OCL) ; Free Art Licence; BSD License.

7. What is personal data; how does one work with it

Solution:

- Personal data are any information relating to an identified or identifiable person
- has a data subject as well as author
- Subject to different laws as well as copyright and/or database right
- Falls under Data Protection Act ⇒ Beware of your responsibilities when using
- To work with personal data:
  - Need consent from data subject
  - Anonymise the data

14 Data infrastructure hardware: filesystems, HSM, data intensive computing

1. *Exam 8.12.2014 Q 3*

(a) What do you understand by the term Amdahl-balanced computer?
Three numbers are used to determine the Amdahl balance of a computer: the Amdahl Number; the Amdahl Memory Ratio; and the Amdahl IOPS Ratio. Define each of these terms.

(b) You are exploring building a system for data-intensive applications and have two possible server blades to evaluate.

i. A 4 core, 2.7 GHz processor blade, with 4 GB total memory and 2 250 GB hard disk drives, each with its own controller, i/o bandwidth of 250 MB/s and each capable of 150 random i/o operations per second.

ii. A 2 core, 1.6 GHz processor blade, with 4 GB total memory and 2 250 GB solid state disks, each with its own controller, i/o bandwidth of 250 MB/s and each capable of 10,000 random i/o operations per second.

Calculate the three Amdahl numbers for each server blade.
Which would you choose for your data-intensive system? Why?

(c) Average IOPS for hard disk drives is often calculated as:

$$\text{IOPS} = \frac{1}{\text{seek time} + \text{latency}} \tag{1}$$

where seek time and latency are measured values quoted in manufacturers specifications. For a hard disk with

$$\begin{aligned} \text{seek time} &= 0.5 \pm 0.1\text{ms} \\ \text{latency} &= 3.0 \pm 0.2\text{ms} \end{aligned} \tag{2}$$

calculate the IOPS rate, including an estimated error.

Solution:

(a) An Amdahl-balanced computer is a computer that has balance between computational power and I/O for use in data-intensive research. The Amdahl numbers are:

- Amdahl number : $\frac{\text{IO bandwidth (baud)}}{\text{CPU Clock rate (Hz)}}$
- Memory Ratio : $\frac{\text{Memory size(bytes)}}{\text{instructions per second}}$
- IOPS Ratio : $\frac{\text{IOPS}\times 50{,}000}{\text{CPU Clock Rate}}$.

For the exam: In multi-core systems, effective CPU clock is CPU clock times number of cores. In multi-disk systems multiply IO bandwidth and IOPS with number of disks if each has its own controller. We assume one instructions per clock (times cores).

(b) Waiting for clarification from Adam.
Possible solution:

i.

$$\text{AN} = \frac{\text{BW}}{\text{CPU}} = \frac{2 \times 8 \times 250 \times 10^8}{4 \times 2.7 \times 10^9} = \frac{4 \times 10^9}{10.8 \times 10^9} = \frac{4}{10.8} = 0.37$$

$$\text{MR} = \frac{\text{MEM}}{\text{IPS}} = \frac{4 \times 10^9}{2.7 \times 10^9} = \frac{4}{10.8} = 0.37$$

$$\text{IR} = \frac{\text{IOPS} \times 5 \times 10^4}{\text{CPU}} = \frac{1.5 \times 10^2 \times 5 \times 10^4}{10.8 \times 10^9} = 0.00139$$

CPU (Hz)	Mem (B)	BW (bs^{-1})	IOPS (s^{-1}) $\times 5 \times 10^4$	AN	MR	IR
2.7× 4	4.0	4.0	7.5×10^{-2}	0.37	0.37	13.9×10^{-}3

ii.

$$\text{AN} = \frac{\text{BW}}{\text{CPU}} = \frac{2 \times 8 \times 250 \times 10^8}{2 \times 1.6 \times 10^9} = \frac{4 \times 10^9}{3.2 \times 10^9} = \frac{4}{3.2} = 1.25$$
$$\text{MR} = \frac{\text{MEM}}{\text{IPS}} = \frac{4 \times 10^9}{1.6 \times 10^9} = \frac{4}{3.2} = 1.25$$
$$\text{IR} = \frac{\text{IOPS} \times 5 \times 10^4}{\text{CPU}} = \frac{10^4 \times 5 \times 10^4}{1.6 \times 10^9} = \frac{5 \times 10^8}{1.6 \times 10^9} = 0.3125$$

CPU (Hz)	Mem (B)	BW (bs^{-1})	IOPS (s^{-1}) $\times 5 \times 10^4$	AN	MR	IR
1.6×2	4.0	4.0	5	1.250	1.250	3.125

(c) To propagate error we use the following formula

$$R\,(X, Y...) \Rightarrow \Delta R = \sqrt{\left(\frac{\partial R}{\partial X}\Delta X\right)^2 + \left(\frac{\partial R}{\partial Y}\Delta Y\right)^2 + ...} \tag{3}$$

Then, say

$$R = \frac{1}{l+s} \tag{4}$$
$$\Rightarrow \frac{\partial R}{\partial s} = \frac{\partial R}{\partial l} = \frac{-1}{(l+s)^2} = \frac{-1}{[\{0.3 + 0.5\} \times 10^{-3}]^2} = -81,632.653$$
$$\Rightarrow \Delta R = \sqrt{[-81,632.653 \times (0.1 \times 10^{-3})]^2 + [-81,632.653 \times (0.2 \times 10^{-3})]^2}$$
$$\Rightarrow \Delta R = \sqrt{66.64 + 266.55} \approx 18$$

We also have

$$R = \frac{1}{s+l} = \frac{1}{(0.5 \times 10^{-3}) + (3.0 \times 10^{-3})} = \frac{1}{3.5 \times 10^{-3}} \approx 286$$
$$\therefore IOPS = 286 \pm 18Hz \tag{5}$$

Helpful: $\frac{d}{dx}x^n = n * x^{n-1} \Rightarrow \frac{d}{ds}\frac{1}{s+l} = -1\frac{1}{s+l}^2$ and due symetry the same result for $\frac{d}{dl}$.

2. What I/O elements are used in modern systems? Give their approximate relative performance for random and sequential operations.

Solution: Large scale computer systems have a mix of i/o hardware.

HDD Hard disk drives

- Spinning, mechanical.
- Access times governed by head seek times
- Performance partly dependent on controller interface: USB, IDE, SCSI/SAS, SATA
- Frequently connected in arrays for resilience & redundancy (RAID)
- Performance of HDD will be used as baseline for the next types

SSD Solid state disks

- Non-volatile flash memory arrays
- Wrote lifetime limited: capacity degrades with each overwrite
- Performance 2x for sequential read
- Performance 50x for random read

LTO linear tape - open

- Most common (open) format for modern tape systems
- Performance comparable to HDD for sequential operations
- Performance terrible for random access

3. Explain hierarchical storage management

Solution: HSM uses policy based software to move data automatically (behind the scenes between tiers of storage. May be for backup, archive, staging.

Performance tier SCSI Raid, SSD

Capacity tier SATA Raid

Archive tier Optical disks, Tape, SATA/IDE/USB disks, cloud

Policies may relate to

- Time since last access
- Fixed time
- Events

4. Which file systems are used for high performance applications?

Solution:

GPFS IBM's General Parallel File System

- Software defined storage layer
- single file system namespace across distributed storage
- Can use any storage hardware underneath
- supports flash based storage of file system meta data
- Parallel file access well-suited to HPC

Lustre Open source parallel file system

- Similar to GPFS
- "a bit faster but less reliable"

HDFS Hadoop file system

- Java based middleware layer for distributing data in chunks across cluster storage
- underpins Hadoop/MapReduce type computational patterns
- Combines distributed file access with redundancy

5. What is data intensive computing? What infrastructure is used?

Solution: Computing applications which devote most of their execution time to computational requirements are deemed compute- intensive and typically require small volumes of data, whereas computing applications which require large volumes of data and devote most of their processing time to I/O and manipulation of data are deemed data-intensive. Wikipedia

Traditionally, data is fed to be computed. In the future, computations are expected to be brought to data.

6. What is a data scope?

Solution: Finding patterns in large amounts of data, data mining, discovering what you're looking fore when collecting data

15 Distributed authentication & authorisation in data grids

1. Explain the difference between authentication and authorisation.

Solution: Authentication (AuthN) is proving identity. Are you who you claim to be?

Authorisation (AuthZ) is checking whether you are allowed to do what you want to do. Can you access this file?

2. What are digital certificates and how do they work? How are they issued? What are they used for?

Solution:

(a) A digital certificate is an authoritative assertion of an association between a public key and an identity.

(b)
   i. Key pair is generated - one public, one private.
   ii. Certificate signining request (CSR) is issued
   iii. CSR is submitted to a Certification Authorities (CA).
   iv. CA "blesses" and issues the certificate.

(c) They are put into use whenever a user wishes to confirm his identity and assert that he is authorise to access certain data.

3. What do we mean when we say public key and private key?

Solution: A cryptographic key is a string of characters that serves as input into an encryption algorithm. It consists of two halves; public and private.

A public key may be shared with anyone and is used for the public to confirm your identity for encrypted documents.

A private key is secret and must be protected. It is used for the algorithm to confirm your identity when decrypting documents.

4. Briefly explain the concept of asymetrically encrypting data.

Solution: In asymetric cryptography every user has two keys: A public key that can be shared and a private key that must be kept secret. There is a encryption function $E(x, K_1)$ and a decryption function $D(x, K_2)$ with input x, public key K_1 and private key K_2. Secure messages are possible because $E(D(x)) = x = D(E(x), x)$ and D can not be reconstructed from E.

Example: Alice wants to send Bob a secret message. They exchanged their public keys at a crypto party. Alice encrypts her message using Bob's public key and sends the generated cipher. Bob can decrypt it using his private key and reply using the same concept encrypting his message with Alice's public key.

5. Draw a diagram of a login sequence. Briefly describe each of the concepts used.

Solution: User is connected to: Identity Provider (IDP), Service Provider (SP), location (WAYF).

IDP: A service that provides your identity, such as a home institution in order to authenticate a user.

SP: Service running at the service you want to access (say, Nature, Elsevier). Makes decisions for the user depending on thier attributes. Decides what parts of identity to accept.

WAYF: Tells the service provider which login endpoint the user will be redirected to.

16 Future challenges: scale, economics, how long term is long-term?

1. How long should scientific data be preserved? What are the economics?

Solution: Scientific data should be preserved as long as it is pertinent. Storing this data electronically requires expenditure towards storage hardware, electricity, and internet.

2. What is Kryder's Law? What is Moore's Law?

Solution: Kryder's law applies to disks; magnetic domain density increases over time. Between 1990 and 2005, disk capacity has increased 1000-fold. This postulate is not sustainable since this has been true since 1980, but it has stopped being true since 2010.

Moore's Law applies to CPUs; every 18 months, transistor density doubles.

3. What are alternatives to hard drives for long term storage?

Solution: Linear tape, Optical storage media, solid state, cloud storage.

4. What are the pros and cons of storing data in the cloud?

Solution:

Pros: Hardware stored in server farms, maintained by someone else, inexpensive.

Cons: Consist of the same hardware as offline storage (optical, magnetic, solid state), constrained by internet bandwidth (not suitable for moving around very large files), you are **no longer in control of your data** - this means that the cloud storage provider does not guarantee safety for your data. If your data is lost, they may take responsibility, but it is unknown what they may do about it.