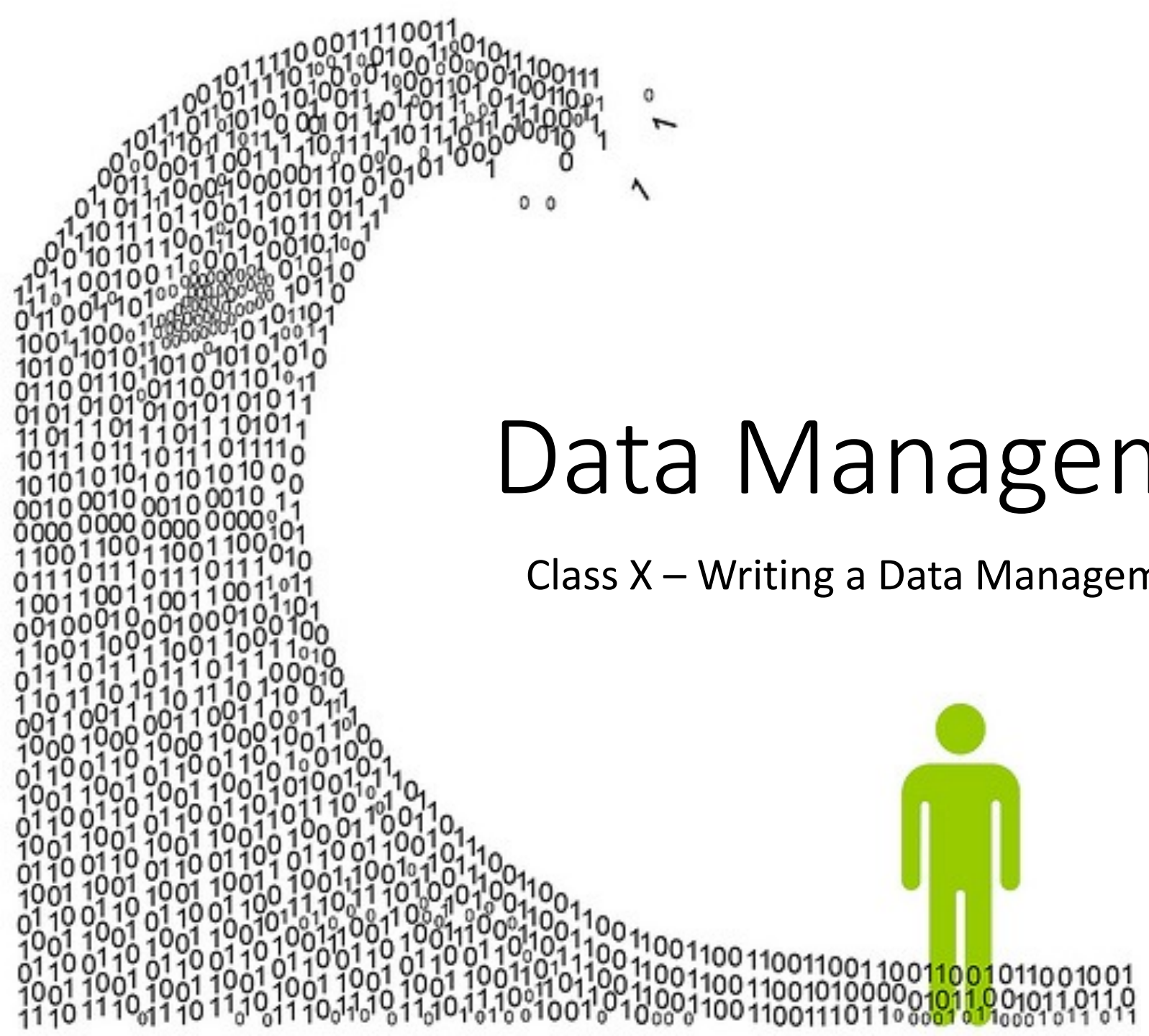
Data Management

Class X – Writing a Data Management Plan



What is a Data Management Plan (DMP)?

- A Data Management Plan is a document produced as part of a research proposal.
- It describes the data you plan to create or acquire during your project; how it will be stored, managed and analysed; and what will be done at the end of the project to preserve the data and ensure it can be shared with other researchers.

When do you need a DMP?

- It’s increasingly common for DMPs to be requested by funding bodies as part of the application process for a research grant.
- Many bodies which distribute public funds (like the European Research Council) have rules which say that research data must be shared publicly. Grant applications need to demonstrate a DMP which will allow this to happen.
- Individual universities sometimes have even stricter requirements for DMPs than the funding bodies – for example, insisting that any budget for funding must include IT costs related to data management.

What to include in a DMP

- *Various institutions have different rules and formats for DMPs, but in general you'll need to include the following details:*
- What data you will generate, collect, or process;
- How that data will be handled and managed:
  - Where and how it will be stored;
  - How it will be backed up and secured;
  - Who will have access to it, and how;
- Whether the data will be shared / made open access;
- How the data will be handled (i.e. archived, preserved and shared) *after* the research project.

Other points to include…

- Does your research project include sensitive data? How will this be handled?
- Similarly, will you be using commercial data? Who will have the license to handle this, and how will it be dealt with?
- Are there extra costs involved in data management for your project?
  - For very large projects with high costs, you may also need to consult your university or funding organisation's rules for awarding contracts – there could be an official process you have to go through to select something like a Cloud platform provider.

What not to include in a DMP

- A DMP is a management document, not a technical document.
- Don’t include technical details of how you’ll manage your data – e.g. SQL table descriptions or query commands, pieces of programming code, or exact system specifications.
  - You may, however, want to include details of which members of your team actually have the technical skills to accomplish the things you’re describing.
- Similarly, you don’t have to get too detailed about your analysis – that’s part of the main research plan. Just give enough detail to show that your DMP is suited to this kind of analysis.

DMP Checklist

- Let’s go through each of those points individually and look at the questions your DMP should be answering.

1) Data Specification

- What data will your project collect?
  - This might include existing data sets you'll use;
  - Data you'll download from the web or from APIs;
  - Data you'll collect through surveys or experiments;
  - Data you'll generate, like labelled training data sets.
- Think about the most suitable format for each type of data.
- Also think about how big each type of data will be. Most projects end up working with a mixture of large (millions of items) and small (thousands of items or less) data sets.

2) Data Processing

- You don’t need to give specific technical details about data processing, but you should mention any major steps you plan to take.
- For example, text analysis will likely include tokenisation steps – remember that the tokenised text is also another kind of data!
- If you’re working with Internet data, you may plan to process it in order to remove bot accounts or spam.

3) Data Storage

- This is really the biggest question from a technical point of view… How will you store your data, and where?
- For each kind of data, you should consider several factors.
  - How big is the data, and what format is it in?
  - Who needs to have access to it?
  - What kind of analysis will be done on it?
  - What level of security does it require?

- This decision should be taken after considering all the other factors in the DMP.

4) Data Sharing (during the project)

- If you're working with a team of people, you need a strategy for sharing data while the project is underway.
- This could be as simple as a shared Dropbox folder, or as complex as a custom database server running on a Cloud platform. Just make sure it's properly suited to your team's needs and requirements.
- Think about your team members' roles.
  - Do they just need to be able to *see* the data?
  - Or do they need to be able to edit and add to it?

5) Security and Privacy

- How will you ensure that your research data is secure (i.e. can only be accessed by authorised people)?
- This is especially important if you are dealing with private, sensitive or commercial information. Any data like that should be highlighted and discussed in detail in the DMP.
- For most projects and types of data, the standard security of password-protected folders on Dropbox or user accounts on a Cloud platform should be enough.
  - But you should still *say* that that's what you're relying on – show that you've at least thought about these issues.

6) Backup Policies

- How will you back up your data?
- How often will you do it?
- Will you back up to a service online, to an external hard drive..?
- If you’re using a cloud service, backups are less of a problem – but it’s still a good idea to keep a copy of your data on a hard drive somewhere secure.
- A fireproof safe in someone’s office with a portable SSD holding your research data is a pretty good investment of a few hundred Euro.

7) Sharing and Access Policies

- Which parts of your data do you plan to share with the public?
- Many funding bodies will insist that you share *all* data that is not private or sensitive.
- Your policy for providing access to your data is in some ways the most important part of the DMP. A good concept to bear in mind is the ERC's "FAIR" policy for research data:
  - Findable
  - Accessible
  - Interoperable
  - Reuseable

“FAIR” Data Access

- **Findable**
  - Your data should be easy to find, so pick a location for it where researchers can search and discover it easily.
- **Accessible**
  - The data should be easy to access. People shouldn’t have to pay to see it, and they shouldn’t have to contact you to get access.
- **Interoperable**
  - The data should be stored in a file format that works with other people’s research software.
- **Reuseable**
  - The data should be in a state where it can be used for future research – either for replicating your study, or for use in a new study.

8) Metadata

- "**Metadata**" is a term that essentially means "data about data".

- For example, every file on your computer *contains* data of some kind- but your computer also holds *metadata* about it, such as the date on which the file was created, which user created it, what application it is for, and when it was last modified.

Research Data Metadata

- In the context of research data management, “metadata” refers to additional data which you create that describes your data.
- For example, you might create keywords and a short description to help researchers to find your data in an archive when they search for your research topic.
- It’s also recommended to create descriptions of your actual data files, so other researchers know what your variables actually represent and how they were collected or calculated.

Including Metadata in a DMP

- It’s recommended to include a short description of the metadata you’ll generate and make available when you release the data from your project.
- Bear in mind that creating metadata is a task someone is going to have to do!
  - Include it in your DMP to show that you’re aware of the necessity for this step, and that you’re planning to ensure it’s carried out correctly.
- You don’t need to go into great detail – just say briefly what kind of metadata you’re going to create and for what purpose.

9) Archiving and Curating

- When your research project is complete, how will you ensure that your data is safely stored?
- This usually involves:
    1. Putting your data files somewhere that is secure and accessible – for example, sites like the Harvard Dataverse, or a journal publisher's data archive, or even somewhere like Github.
    2. Keeping a secure backup copy of the files at your own institution – most universities have a facility for doing this.
    3. Finally, keeping your *own* archive of the files – preferably one that will last for a long time. Burning files to a Blu-Ray disc is a good option for this; the discs can theoretically last hundreds of years.

10) Costs

- This isn’t something every research proposal needs to include, but sometimes it’s listed as a requirement.
- Data management costs could include:
  - Cloud service platform fees (for a big project, these could be thousands of Euro per month, but for a smaller project it’s not uncommon to pay less than €10 per month).
  - Hardware costs – buying and building servers, backup systems etc.
  - Staff costs – if you need to hire an expert to implement part of your data management plan.

- **<u>You don’t need to include cost details in the DMP for your final assessment.</u>**

Using Pre-existing Data Management Plans

- Depending on which service you're going to use to host your public research data long-term, you may be able to use some parts of a pre-existing DMP.
- **For example:**
- ICPSR (*Inter-university Consortium for Political and Social Research*) is a group of more than 750 universities and research organisations who promote good data management, analysis and archiving practices in political and social sciences.
- It maintains an archive of research data hosted by the University of Michigan: www.icpsr.umich.edu

IPCSR's Data Management Plan

- IPCSR has a document which outlines the steps it'll take with your research data archive to ensure it's kept safe, secure and accessible.
    - A copy of it is included in the Files package for today's class, as reference.

- However, this **only** covers management of your data post-research (so you'll still need to write a plan for all your management strategies during the project itself), and you'll need to modify parts of it to fit with the specifics of your project.

- In other words, these pre-written plans are only a starting point.

Other Resources about DMPs

- A great starting point for accessing information about writing Data Management Plans for research is the Stanford Library page on this topic: https://library.stanford.edu/research/data-management-services/data-management-plans

- The European Research Council's guidelines on DMPs for Horizon2020 projects are also useful. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
    - A copy of the ERC's suggested template for DMPs is also included in the files for today's lecture – if you wish, you may use this as a basis for your final assignment submission. It's up to you.

Final Assignment *(25%, due March 25th)*

- Your final assessment should be at least two pages long, and should try to cover all of the points listed here (except costs, which you may omit).
  - Even if you don't think a certain point is relevant to your project, at least show that you've thought about it.

- Remember: there are no real right or wrong answers. Justify the answers you've given and show why you made your choices.
- A DMP is a "live" document – it's expected to change and evolve as the project continues and encounters new challenges!