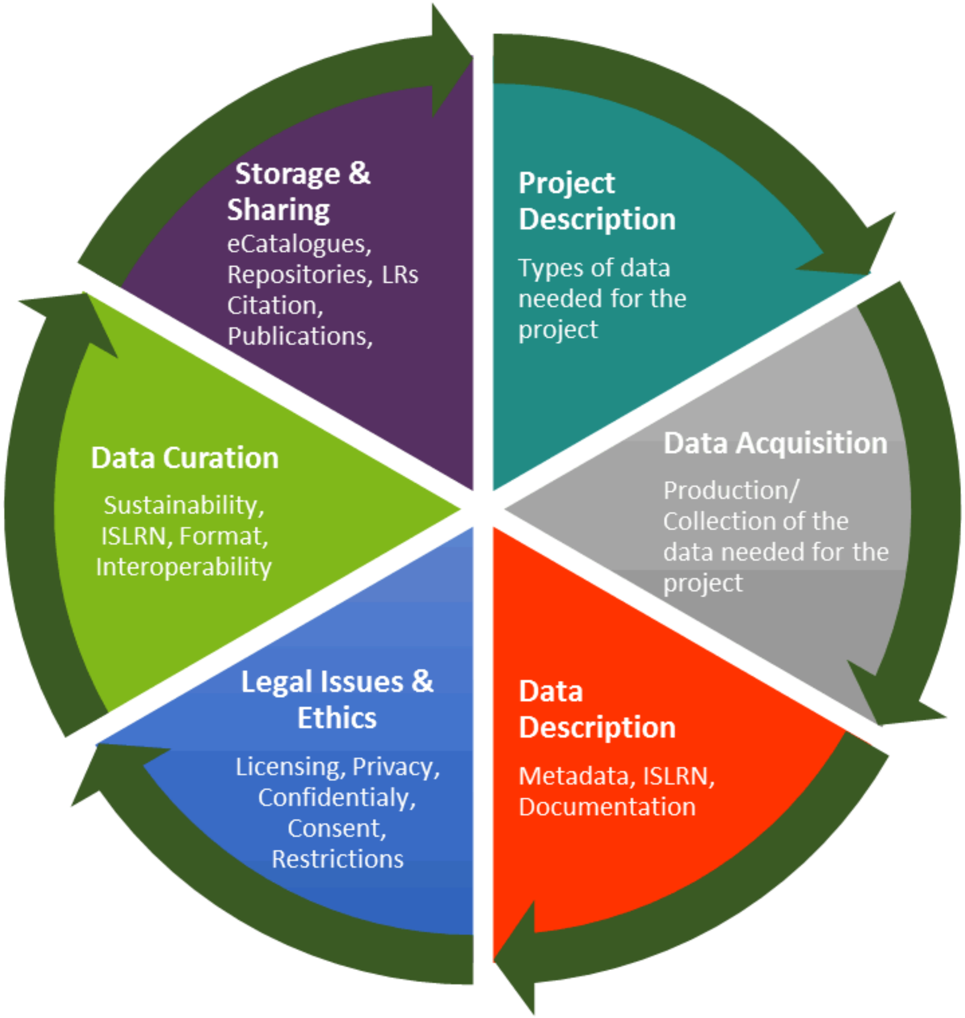
ELRC Workshop in the Netherlands
Identifying and managing your data: Q&A

Jan Odijk / Steven Krauwer
Utrecht University

Data Management Plan (DMP)

The **DMP** outlines how data shall be handled during the production workflow and after. It covers the entire data lifecycle and it defines a **data policy** to manage data efficiently and ensure data is sustainable.


Project Description
Types of data needed for the project
Data Acquisition
Production/ Collection of the data needed for the project
Data Description
Metadata, ISLRN, Documentation
Legal Issues & Ethics
Licensing, Privacy, Confidentialy, Consent, Restrictions
Data Curation
Sustainability, ISLRN, Format, Interoperability
Storage & Sharing
eCatalogues, Repositories, LRs Citation, Publications,

Concerns in creating a DMP

- **Anticipate all potential legal issues**
  - Ensure that your data IPRs are cleared
  - Ensure that the producing parties adhere to your right “ownership” (e.g. relations with LSP: ensure you keep all rights)
  - Ensure that all produced intermediary documents are yours (e.g. translation memories)
  - Check the privacy issues in advance and plan for anonymization if necessary
- **Define your management plan with respect to the task**
  - This has to account for the main goal (e.g. document writing, doc translation, etc.)
- **Plan for repurposing** (from documentation to LRs)
  - Request data in a usable format (not only PDFs but also TMX/XML/)
  - Make sure that your data uses up-to-date media (no CDs?)
- **Foresee future publication and sharing** as Public Sector Information (PSI)

Questions?

Q&A: Copyright

If a public agency outsources a translation of a text of which it owns the rights, who owns the copyright of the translated version? Can the translation be shared?

It depends on what the outsourcing contract establishes with regard to IPR. Public agencies should make sure that the outsourcing contract grants them the right to freely reuse and share translation memories.

Q&A: Copyright

I have created a corpus of literature texts for my research. Can I donate it to ELRC?

All the texts included in the corpus must be IPR cleared. Some of them, especially old works, may be in the public domain (e.g. if copyright has expired). For the rest, a licence must be obtained from the copyright holders authorising redistribution to third parties.

Q&A: Copyright

I am the owner of the translation, but not the owner of the source text (or vice versa). Can I share the parallel dataset? What are the necessary steps to take?

In order to be able to distribute a parallel dataset, IPR must be cleared both for the source text and for the translation. If the source text (or the translation) is copyrighted a licence must be obtained from the owner of the source text (or the translation) to be able to share it with third parties. The first step is thus to contact the owner of the text to find out whether the text is available under an open licence or if a different licensing agreement has to be negotiated.

Q&A: Copyright

We have compiled some bilingual terminological resources from an existing dataset we have, and that is available from a data center, and combined with some other corpora and dictionaries we have at our disposal. We are not sure about whether we can distribute the newly developed resources under one of the CC licences.

If a new resource is built from several pre-existing resources, IPR has to be cleared for all these resources. The licences should allow redistribution and derivative works. If you are the owner of the resources and you have granted distribution rights to a third party you should make sure that the distribution agreement allows you to distribute the resource through an additional channel under CC licences.

Q&A: Copyright

I created a corpus from Wikipedia texts. Wikipedia is licenced as CC-BY-SA 3.0. What should the licence of the derivative be? CC v3.0 or CC v4.0?

Anyone adapting BY-SA works has to apply to the derivative a licence declared compatible with BY-SA. In the case of BY-SA 3.0 all future versions of BY-SA are compatible, so the derivative could be licenced under CC BY-SA 4.0.

Q&A: Copyright

I have a non-CC dataset. Can I contribute a terminology/ a language model derived out of this?

If the derivative resource contains substantial parts of the original dataset (e.g. long citations, full sentences), then a licence must be obtained from the copyright holder to be able to distribute it. However, if the derivative resource does not contain substantial parts of the original dataset (e.g. it contains only statistical information about number of tokens, type occurrences, collocations, …) the dataset can probably be distributed without obtaining a licence from the owner of the original dataset. This is to be studied on a case-by-case basis.

Q&A: Data protection

What do we do with a dataset which contains personal data?

Not all personal data are to be anonymized.
If you have doubts about how to handle datasets which contain personal data contact the ELRC team. ELRC offers a legal helpdesk as well as an anonymization service for donated data.

Q&A: Data protection

I have a collection of publicly available bilingual documents from my public sector organization, e.g. expressions of interest, calls for tenders etc. They include person names, e.g. names of directors, members of committees. Do they fall under the personal data restrictions? Should they be anonymized?

These are listed as public office activities so it does not belong to the private sphere.

Q&A: Data processing

We have data but we do not have the resources to identify the relevant ones and process them.

ELRC can help you identify the relevant datasets. It also offers language processing services to public administrations (data conversion, tag removal, re-formatting, cleaning, alignment, metadata validation, etc.). On-site assistance is also offered to provide technical assistance. These services are free of charge.

Q&A: Data processing

We have a huge collection of scanned pdfs. Can we request on-site assistance? Shall we get back the machine readable forms (the result of OCR)?

Results of OCR of scanned pdfs differ in quality (depending on the languages, paper quality, etc.). Some may be useful for further processing to obtain machine-readable texts, and can thus be further processed by ELRC to produce parallel corpora. ELRC offers an on-site assistance service.

Q&A: Data suitability

Most of our data are numerical (e.g. National Bank, Statistics office) accompanied by some text. Do you still need them?

ELRC main focus is on textual data. However, if your numerical dataset contains text this can still be useful, especially in the case of bilingual or multilingual text.

Q&A: Data suitability

A sample of our national corpus is available through a different repository, e.g. CLARIN. Should I also contribute it to ELRC?

Only if these are different parts of the corpus (ELRC has access to data sets from Data Centers).