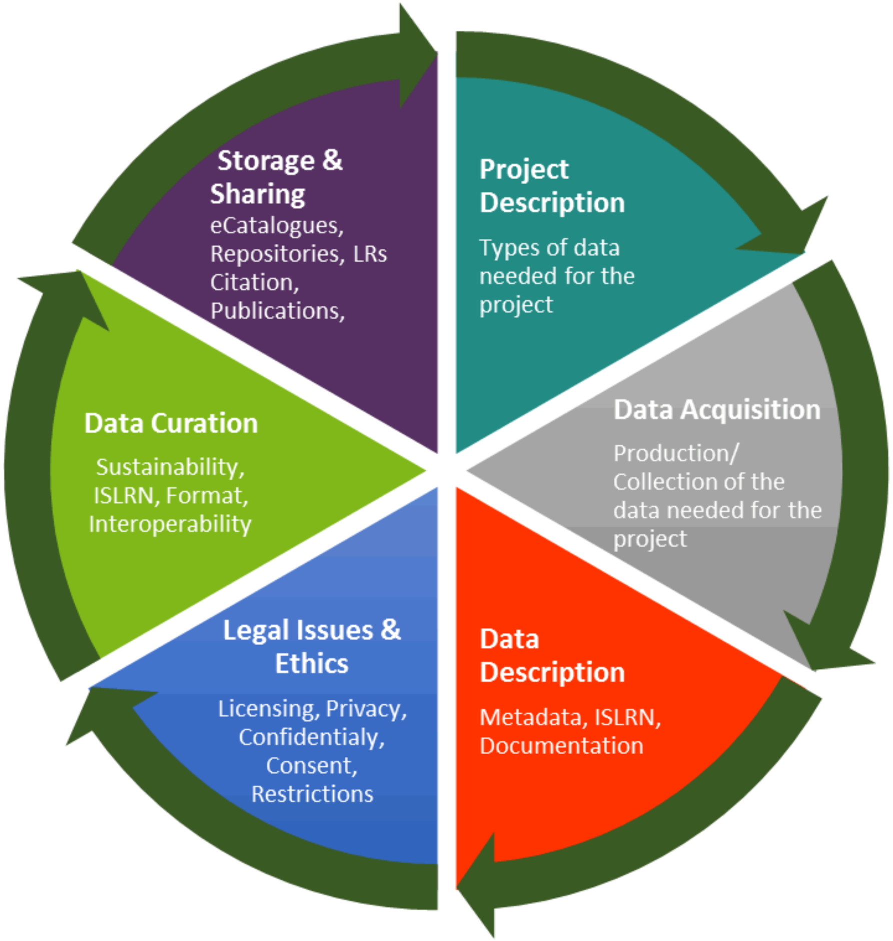
Data Management: Practical Hints and Tips

Khalid Choukri & Pawel Kamocki
(ELDA)

Data Management Plan (DMP)

The **DMP** outlines how data will be handled during a 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.


Storage & Sharing
eCatalogues, Repositories, LRs Citation, Publications,
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

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 medium (no CDs?)
- **Foresee for future publication and sharing** as Public Sector Information (PSI)

Data lifecycle

- Project description
  - Ensure that the original documents are described
  - Ensure that your needs are described
  - Anticipate what you can get as valuable resources (a side effect)
- Data production
  - Whether internal or outsourced, check that the tools used are compatible with your needs and beyond (e.g. CAT, MT, etc.)
  - Ask for the list of tools and production software
  - Check if you can get texts in the multiple languages aligned to each other
  - Keep a clear documentation of the data being produced (meta-data)

Data lifecycle

- Validation
  - In addition to your quality control, you may want to use some of the validation tools (lexical coherence, syntactic analysis, etc.)
- Sharing/distribution
  - Ensure your data falls within the PSI directive as transposed in your country
  - If not, foresee an open and permissive licence
  - If privacy is an issue, plan necessary procedures to handle these
- Maintenance/preservation
  - The best option is often to partnership with a data centre
  - See how ELRC can assist you
  - There is also the “option” of national open data portal
  - Only “putting” data on the web is not a sufficient option (referencing?)

Q&A

- If a public agency outsources a translation, 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 they keep the right to freely reuse and share translation memories. It is also advisable that public agencies
  - It also depends on national legislation (in some Member State may be regulated by law)
- We have data but we do not have the resources to identify relevant data and process them.
  - ELRC offers language processing services to public administrations (data conversion, tag removal, re-formatting, cleaning, alignment, metadata validation, etc.)
- What do we do with a dataset which contains personal data?
  - ELRC can assist in identifying compliance with GDPR
  - ELRC offers also anonymization services.

Outsourcing Translations

- What happens if you outsource (some) of your data production? (e.g. you outsource a translation to a translator)

  - Make sure that you keep the right to freely reuse the translation and share it with third parties (translations are protected by copyright so if there is no copyright transfer from the translator you may not be able to reuse the translations).
  - Outcome of the contract:
    - Translated documents (in your favorite format: .doc, .pdf; etc.)
    - You should also make sure you obtain translation memories in reusable formats (e.g. TMX) as well as terminological data.

Copyright and Translations

- Translations are protected by copyright ‘without prejudice to copyright in the original work’
- Only condition for copyright protection: originality (author’s own intellectual creation)
  - Permission from the author needed to make translations and communicate them to the public
  - Reproduction of translations and their communication to the public require permission from both the translator and the author of the original work

TMX and the Database Right

- *Sui generis* database right (Database Directive 1996)
- Belongs to the maker of the database (person or entity who invested in the production)
- Condition: substantial investment in obtaining/ verification/ presentation of data
- Duration: 15 years after the investment (renewable with new investment)
- Exclusive rights (permission needed to accomplish these acts):
  - Extraction (reproduction) of a substantial part of the database (>10% of the database)
  - Re-utilization (sharing) of a substantial part of the database (>10% of the database)
  - Non-substantial parts can be extracted and re-utilized freely, BUT repeated and systematic extraction and re-utilization of such part require permission

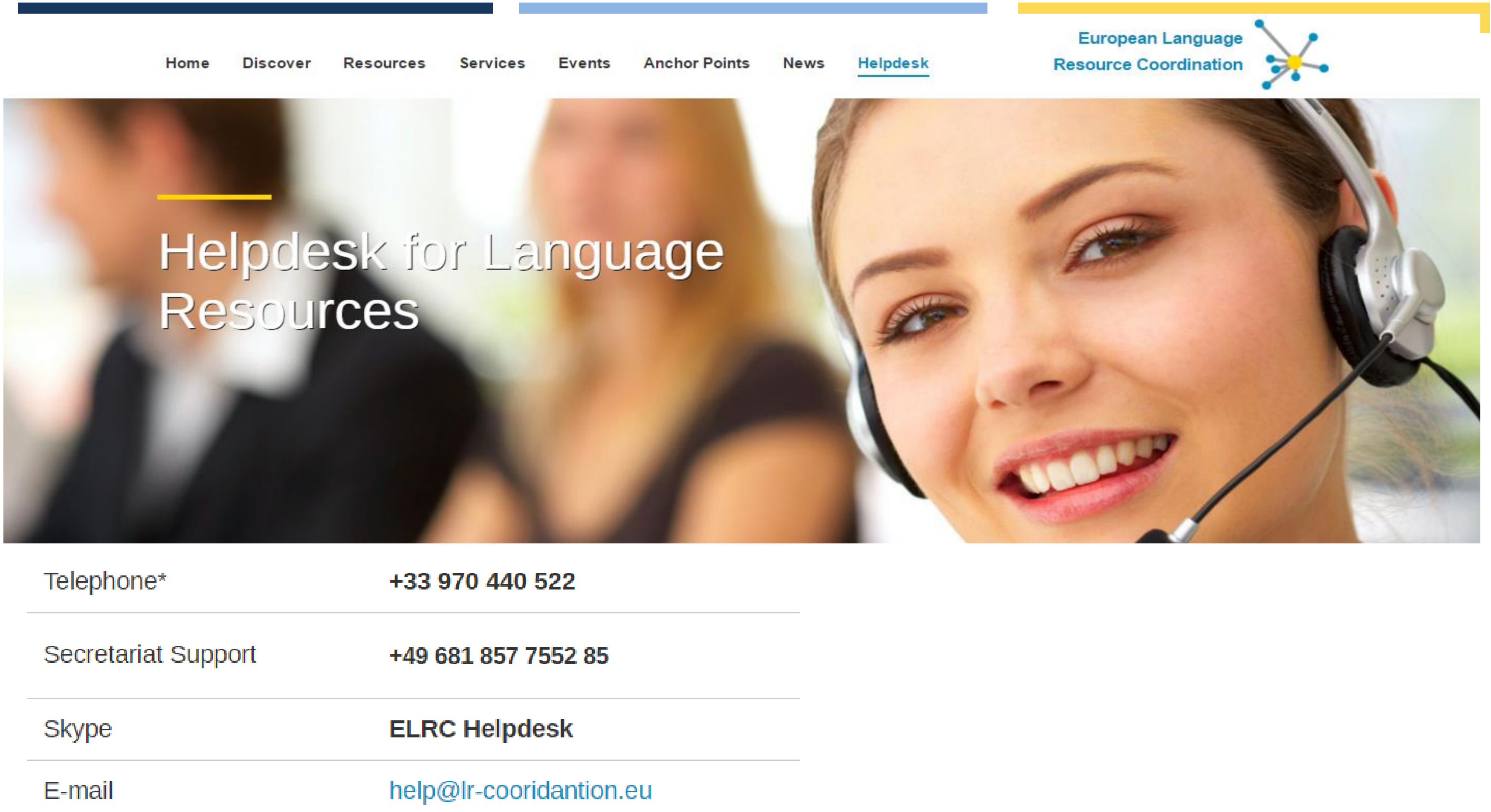
Copyright Ownership in Official Translations

- Depends on the circumstances of the specific case…
- No one can transfer more rights than he has: a license can only be given by someone who holds copyright or at least a sufficiently broad license
- Germany: copyright (incl. in translations) belongs *ab initio* to the author/translator and cannot be transferred
  - A contract for making translations should usually include an exclusive license to use the results (implied in employment contracts, needs to be express in other contracts)
- UK/Ireland: work for hire (copyright belongs ab initio to the employer)
- Other countries (e.g. France): copyright in works of public servants belongs ab initio to the State
- In short: have a look at the contract, esp. with external providers

Copyright and Machine Translations

- Computer-generated works (such as machine translations) are not protected by copyright (they are not 'their authors' own intellectual creations')
- However, computer-assisted creations can be protected by copyright, if the human contribution is original
  - just correcting obvious grammatical mistakes is not enough to claim copyright!
- It is getting more and more difficult to prove that the translation was computer-generated and therefore not protected by copyright
  - some MT tools use watermarking techniques
- The Terms of Use of some MT apps or services (e.g. Google Translate) stipulate that the user grants the provider a licence to use the input data
  - the user loses control over the input!

ELDA Legal Helpdesk



Helpdesk for Language Resources

Telephone*	**+33 970 440 522**
Secretariat Support	**+49 681 857 7552 85**
Skype	**ELRC Helpdesk**
E-mail	help@lr-cooridantion.eu

Webforum: http://helpdesk.lr-coordination.eu/overview/

Thank you for your attention!

Main Concepts in Data Protection

- Personal data:
  - Any information(text/image/audio, fact/opinion, true/false)…
  - …relating to…
  - …an identified (singled out directly or indirectly)…
  - …or identifiable (possible to identify directly or indirectly taking into account any means reasonably likely to be used)…
  - …natural person (no: dead person, legal entity)
- Anonymisation ≠ pseudonymisation
- Processing: any operation performed on personal data
- Controller: person or entity who defines the means and purposes of processing
- Processor: processes data on behalf of the controller