European Language Resource Coordination
Connecting Europe Facility

7th LRB Meeting, 20 September 2018

Legal Session

Pawel Kamocki

Khalid Choukri
ELDA

Valérie Mapelli

European Language Resource Coordination
Connecting Europe Facility

Introduction

- Legal and Ethical Issues in language data collection and re-use
  - Web crawling
  - General Data Protection Regulation
  - Data Management
- Q&A Session
- ELDA Legal Helpdesk
  - Assisted by iRights (Berlin law firm) and FIDAL (Paris law firm)

European Language Resource Coordination
Connecting Europe Facility

Legal Matters in Web Crawling

European Language Resource Coordination
Connecting Europe Facility

Web Crawling

- Web crawler: a piece of software browsing the Internet in a methodic manner, copying web pages that meet pre-defined criteria
  - such copies can be used e.g. to build language resources
- Crawling is subject to legal restrictions
  - Copyright
  - Sui generis database right
  - Digital Rights Management (technological protection measures)
  - Personal Data Protection
- Kamocki, Popescu, *ELRC Report on legal issues in web crawling,* publication imminent

European Language Resource Coordination
Connecting Europe Facility

Copyright and Web Crawling I

- Web content is often subject to copyright protection
  - exceptions: official documents (in some jurisdictions), factual statements (e.g. timetables)
- Principle: copying of copyright-protected content requires permission from the rightholder
  - unless it is permitted by an exception (e.g. research, quotation, temporary acts of reproduction)
- Copyright exceptions are harmonized by the Copyright Directive 2001
  - Temporary acts of reproduction (copies need to be temporary)
  - Research (only non-commercial; often additional requirements at the national level)
  - Private copy (only for purely private, non-professional use)
  - Citation (copies need to be integrated in a larger work)
- None of the current copyright exceptions can cover a broad spectrum of crawling activities

Copyright and Web Crawling II

- New exception for TDM activities (art. 3 and 3a, Directive on Copyright in the Digital Single Market 2019)
  - Requirement: lawful access (expressly authorized or not expressly prohibited?)
  - Beneficiaries: public research organizations (mandatory), general public (optionally)
  - No communication to the public (sharing) allowed (citations only)
  - No repurposing
  - Not overridable by contractual clauses, but still overridable by technological protection measures (in theory: only to protect the security and integrity of networks and databases…)
- Exceptions for TDM for non-commercial research exist already in some Member States (Germany, France)
- Significant relief for public research organizations, but far from being completely satisfying for the general public

Roadmap for Web Crawlers

- *A priori* source clearance
- Are the contents available via the URLs:
  - Protected by copyright and/or related rights?
  - Available under a public license (such as Creative Commons) or other permissive conditions?
  - Public Sector Information (documents held by public bodies)?
- 1st question answered in the negative OR 2nd or 3rd question answered in the positive => content can be lawfully crawled if it does not contain personal data

European Language Resource Coordination
Connecting Europe Facility

GDPR and How It Affects Our Work

GDPR

- The General Data Protection Regulation
  - Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data
  - Adopted on 27 April 2016
  - Applicable on 25 May 2018
- Replaced the Personal Data Directive
  - Directive requires implementation (choice of forms and means left to the Member States)
  - Regulation is directly applicable in all Member States

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

European Language Resource Coordination
Connecting Europe Facility

GDPR in a nutshell

- Lawfulness, fairness and transparency
- Purpose limitation
- Data minimisation
- Data accuracy
- Storage limitation
- Integrity and confidentiality
- Accountability

- Data protection by design and by default
- Data protection impact assessment
- Fines up to 20 000 000 EUR or 4% of total annual turnover

Data Transfers under the GDPR

- Within the EU: free (GDPR principles need to be respected)

- To third countries: forbidden, unless:
  - EC issued an adequacy decision (not for the US!) OR
  - Respect of the GDPR is part of a contract or binding corporate rules OR
  - Data subject has given his explicit consent, after being informed about the risks

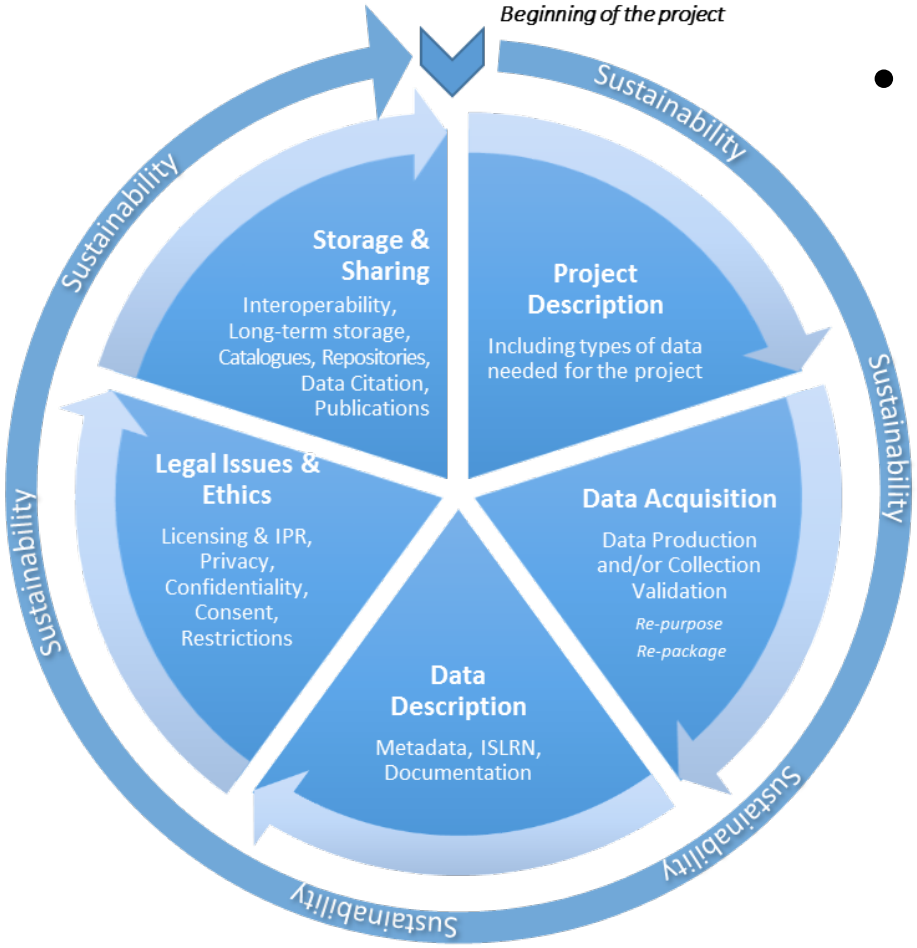
Processing for research purposes

- Possible without consent (legal ground: legitimate interests)
- Data collected for a different purpose can be re-used for research purposes (purpose extension)
- Member States may provide for derogations from certain rights of data subject (access, rectification, restriction, objection)
- Condition: *appropriate safeguards*
  - e.g. enhanced transparency (a publicly available record of processing)
  - enhanced security (state-of-the-art encryption methods)
  - consent, impact assessment – even if not required

European Language Resource Coordination
Connecting Europe Facility

Data Management – Contracting LSPs

Processing for research purposes



- Data Management Plan
  - Since 2015, ELRA has proposed a DMP for language data
  - 2017: extension of the H2020 Open Data Pilot (‘open by default’, obligation to adopt a DMP)
  - US National Science Foundation requires a DMP for all grant proposals

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

European Language Resource Coordination
Connecting Europe Facility

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

European Language Resource Coordination
Connecting Europe Facility

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!

European Language Resource Coordination
Connecting Europe Facility

ELDA Legal Helpdesk

Home Discover Resources Services Events Anchor Points News Helpdesk

European Language Resource Coordination

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/

European Language Resource Coordination
Connecting Europe Facility

Thank you for your attention!