European Language Resource Coordination
Connecting Europe Facility

"Getting Data and Language Resources: Technical & Practical Issues"

Data Management Basic Workflow

Dr. Khalid Choukri
(Evaluation & Language resource Distribution Agency)

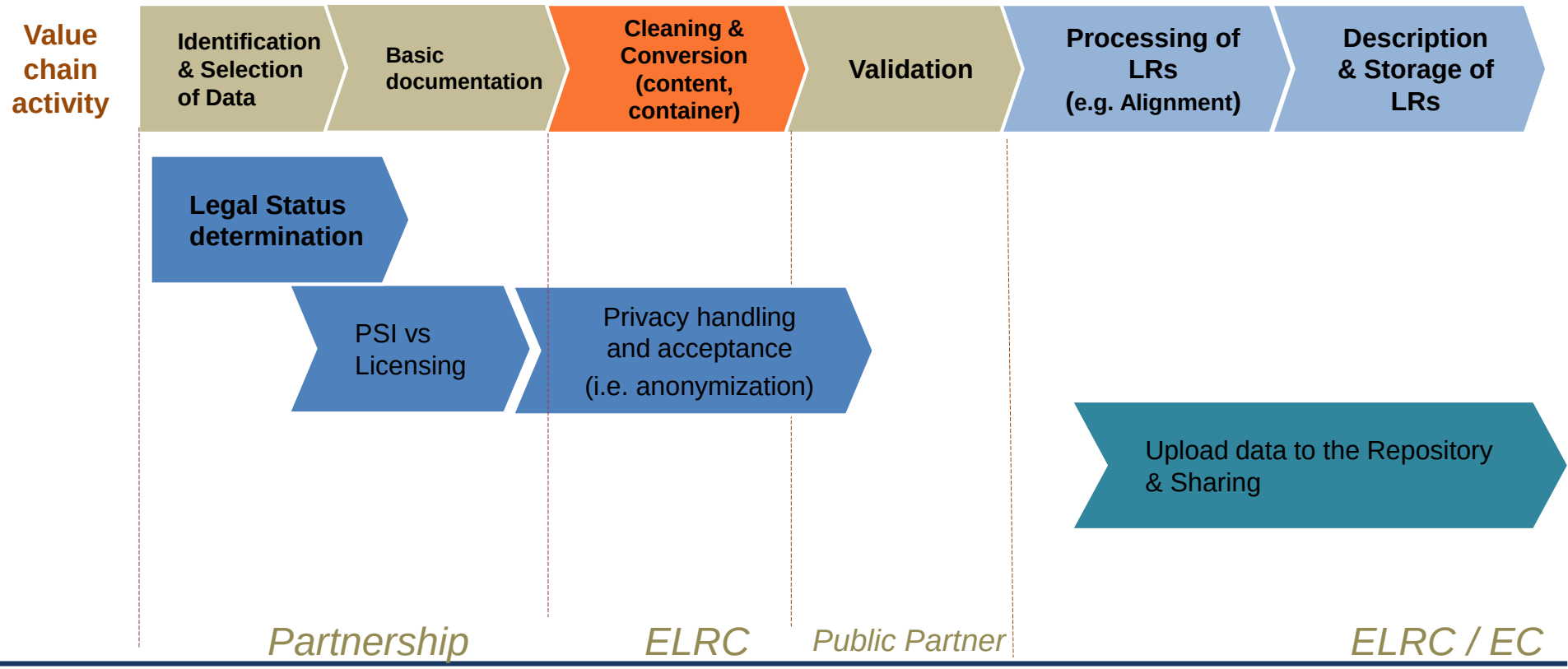
European Language Resource Coordination
Connecting Europe Facility

Overview

- Please refer to Session 7A for “What data & Why”
- Data-driven Paradigm
- Data is needed in all language(s)
- Where can we discover Data: Public Sector Players
  - Visible data e.g. Web (HTML pages, reports, etc.)
  - Invisible Data: archives , hidden web, internal repositories
  - Through Language Service Providers

European Language Resource Coordination
Connecting Europe Facility

Illustration of Data Packaging Workflow Data ➔LRs (Language Resources)



Issues to address (1)

- Issues to start with:
  - Identification of sources, identification and selection of data sets (raw data)
  - Privacy and ethics management from the start (e.g., anonymization when required, accept/reject data)
  - Documentation with basic identification elements (languages, domains, year, …)
  - Choice of medium and data formats for the transfer of the “raw” data (preference for the ELRC ad hoc platform)

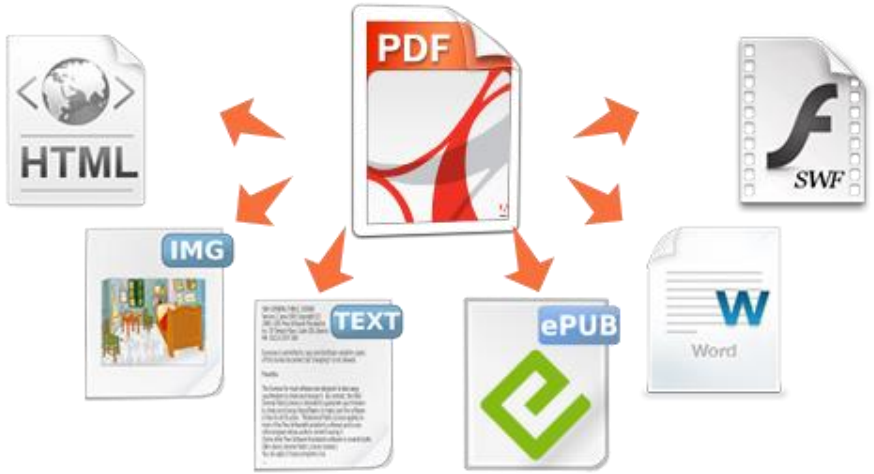
European Language Resource Coordination
Connecting Europe Facility

Issues to address (2)

- Technical issues at the preparation/packaging stage:
  - Cleaning of data format (discarding formatting, encoding character sets e.g. UTF8, formatting features e.g. bold, italic; graphics, ads, tables, html tags, etc.)
  - File cleaning (e.g. conversion to XML, XLIFF, etc.)
  - Data preparation for Automated Translation tools (e.g. alignment)
  - Validation and Quality Control of the output (Language Resource format, content, storage)
  - Description of the Language Resource (meta-data)
  - Packaging and delivery (data repository with e-sharing) to EC and Owner

European Language Resource Coordination
Connecting Europe Facility

Any Digital Textual Data !!



Formatting Example

Greece is a place of culture, the arts and sciences. Its tradition of contribution to global cultural and scientific communities, combined with its outstanding natural beauty and **excellent infrastructure**, has made it an ideal place in which to hold conferences. Over the last few years, Greece has more and more frequently welcon
have participated
International Airp
airports in the wo
organization of int

Greece is a place of culture, the arts and sciences. Its tradition of contribution to global cultural and scientific communities, combined with its outstanding natural beauty and excellent infrastructure, has made it an ideal place in which to hold conferences. Over the last few years, Greece has more and more frequently welcomed people of letters, sciences and the arts, who have participated in symposia, conferences and exhibitions. Athens International Airport 'Eleftherios Venizelos', one of the most modern airports in the world in operation since 2001, greatly boosted the organization of international conferences.

Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. Η μακραίωνη συμβολή της στο παγκόσμιο γίγνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις **άρτιες υποδομές**, την καθιστούν ιδανικό τόπο
ελληνική
ώπους των
οποίοι
σεις. Ο Διεθνής
ένα από τα
ποίος
στη

Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. Η μακραίωνη συμβολή της στο παγκόσμιο γίγνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις άρτιες υποδομές, την καθιστούν ιδανικό τόπο διεξαγωγής συνεδρίων. Τα τελευταία χρόνια, η ελληνική επικράτεια υποδέχεται όλο και συχνότερα ανθρώπους των γραμμάτων, των επιστημών και των τεχνών, οι οποίοι συμμετέχουν σε συμπόσια, συνέδρια και εκθέσεις. Ο Διεθνής Αερολιμένας Αθηνών «Ελευθέριος Βενιζέλος», ένα από τα πλέον σύγχρονα αεροδρόμια παγκοσμίως, ο οποίος λειτουργεί από το 2001, έδωσε μεγάλη ώθηση στη διοργάνωση διεθνών συνεδρίων.

European Language Resource Coordination
Connecting Europe Facility

Issues to address (3)

- Legal Issues
  - Legal status determination (accept/reject decision)
  - Accuracy and acceptance of privacy processing (e.g., anonymization)
  - Application of PSI versus need for a License
- Practical issues
  - Role of the ELRC Consortium (technical/legal helpdesk, repository, consultancy)

European Language Resource Coordination
Connecting Europe Facility

Cooperation Actions

- Identification of sources, identification and selection of data sets (raw data)
  - Data can be obtained from the visible sources (e.g. harvested from web)
  - Data can be handed over by the public sector players
  - Public sector players can boost the identification of visible sources
- Processing described above can be carried out in cooperation by the ELRC and the data provider

Legal elements

- Procedural Issues (data requests vs. open by default e.g. PSI)
- Licensing
  - ELRC can help with the procedures
  - Model licensing agreements
    - Government Open Licenses
    - Standard Re-use Licenses
    - License interoperability

European Language Resource Coordination
Connecting Europe Facility

We need your involvement

- You know your data
  - visible vs. invisible
- Access to archives, deep web, etc. is often not possible from the outside.
- Not all data is already under PSI or a permissive license
- Access to derived forms (e.g., PDF) is less efficient than access to internal source content repositories.

European Language Resource Coordination
Connecting Europe Facility

Data anonymization

- Identify a large source of data on individuals, organizations etc.
- Use a Named Entity Recognizer (NER) to find and remove private bio-data (names, locations, dates, birth information, etc.) and replace with generic placeholders.
- Confirm results meet acceptable requirements
  - Reject data if anonymization not accurate as required

European Language Resource Coordination
Connecting Europe Facility

Conclusions

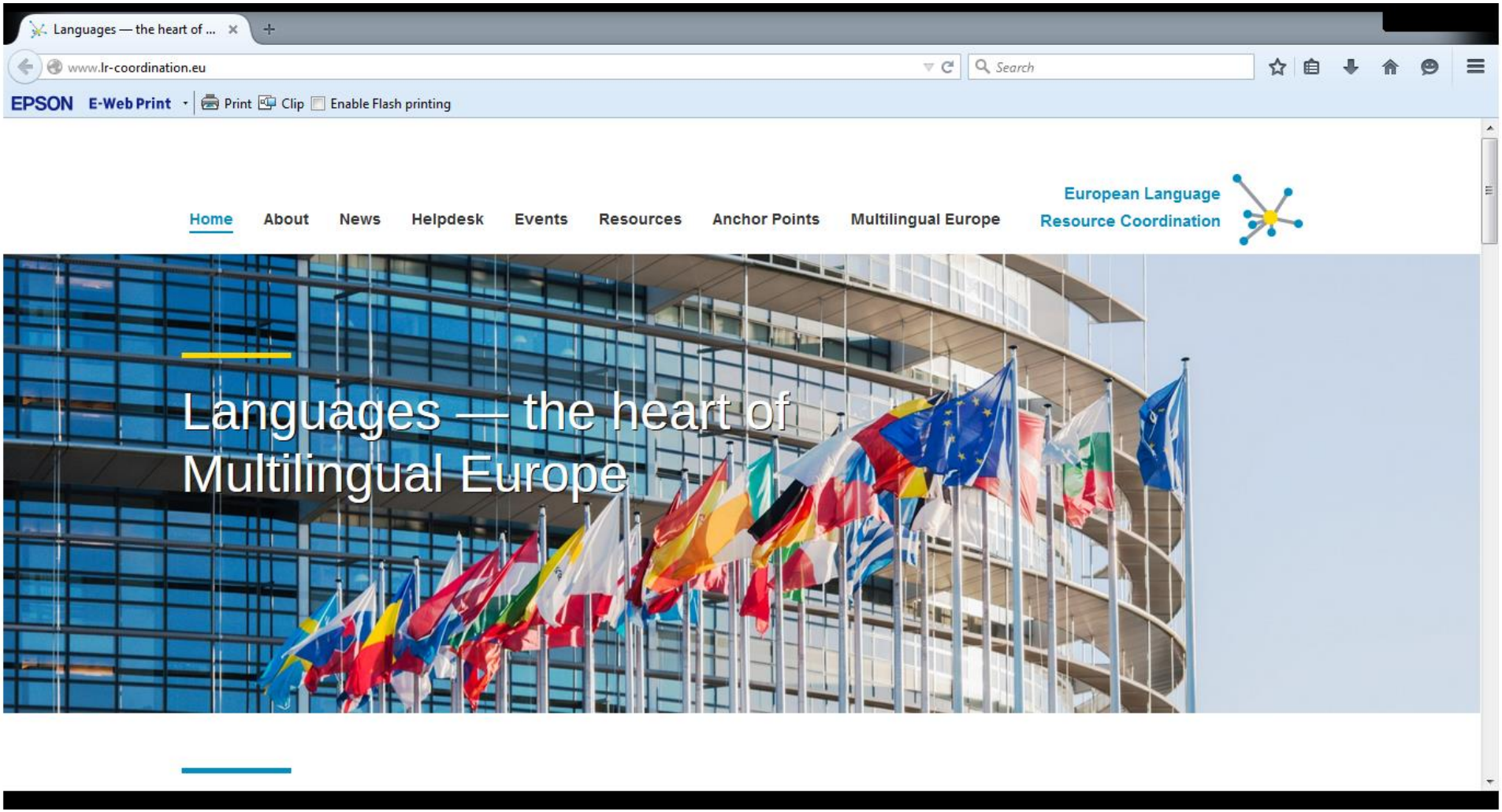
- Repurposing existing data (human translations) is the best way to improve Automated Translation quality.
- Data-driven paradigms provide an efficient way to leverage value from existing resources.
- ELRC can help reviewing data for suitability (at any phase)
- Do not underestimate the value of your language resources, foresee a Data Management Plan

European Language Resource Coordination
Connecting Europe Facility

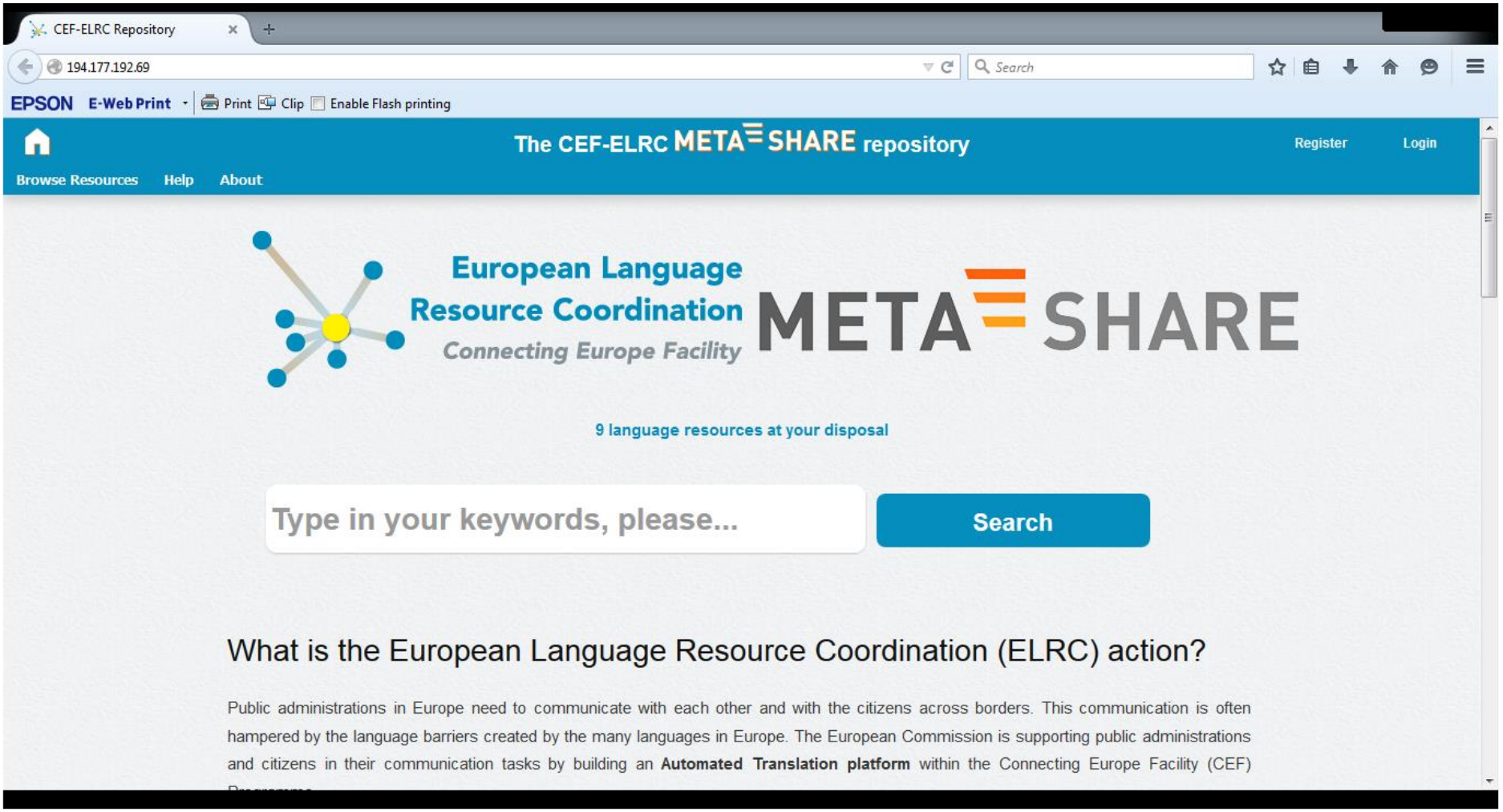
Helpdesk and Support

European Language Resource Coordination
Connecting Europe Facility

ELRC Portal

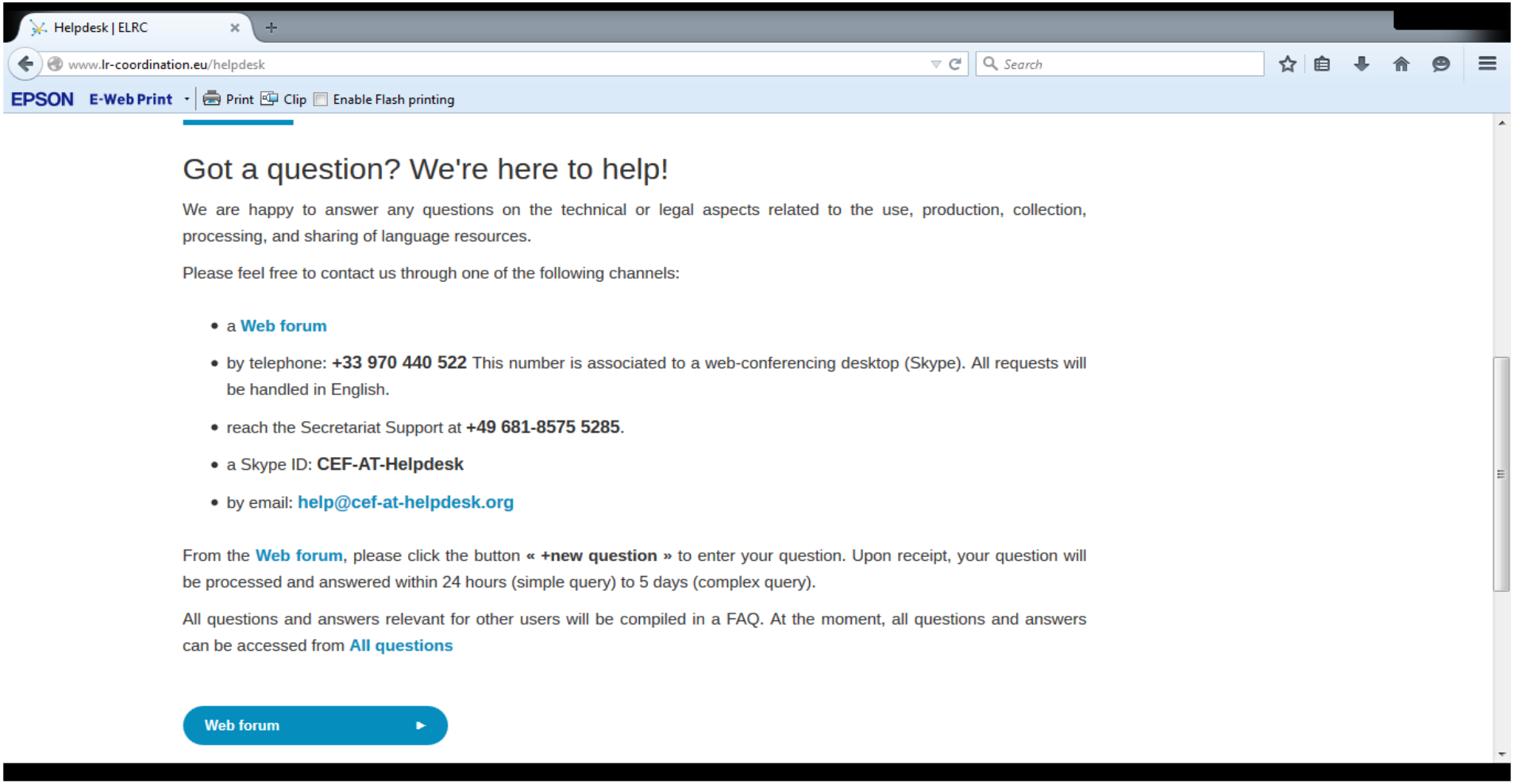


ELRC Portal: Repository



European Language Resource Coordination
Connecting Europe Facility

ELRC Portal: Helpdesk



ELRC Portal: Web Forum