

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

Data Management Basic Workflow

Dr. Khalid Choukri

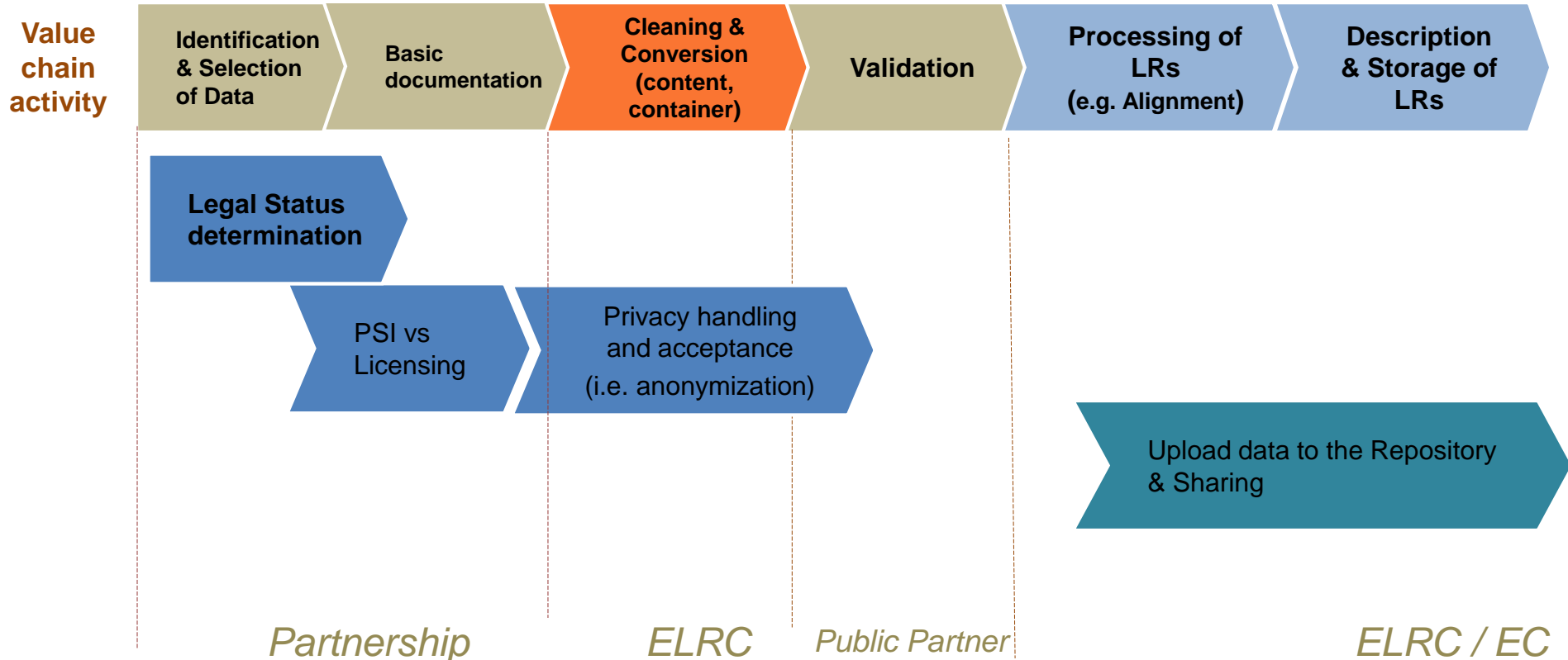
(Evaluation & Language resource Distribution Agency)



- 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

Illustration of Data Packaging Workflow

Data → LR (Language Resources)





- 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)

- 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

Any Digital Textual Data !!





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.

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 γίνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις **άρτιες υποδομές**, την καθιστούν ιδανικό τόπο διεξαγωγής συνεδρίων. Τα τελευταία χρόνια, η ελληνική

Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. Η μακραίωνη συμβολή της στο παγκόσμιο γίνεσθαι, 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.

ώπους των οποίων ο οποίος είναι από τα ποίος στη



- 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)

- 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



- 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

We need your involvement



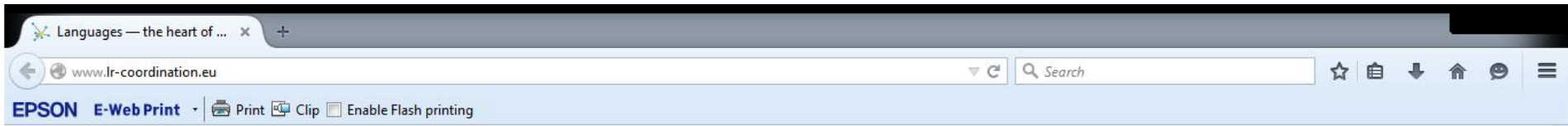
- You know your data
 - visible vs. invisible
- Access to archives, deep web, etc. is often not possible from the outside.
- Not all data is already under PSI or a permissive license
- Access to derived forms (e.g., PDF) is less efficient than access to internal source content repositories.

- 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



- 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

Helpdesk and Support

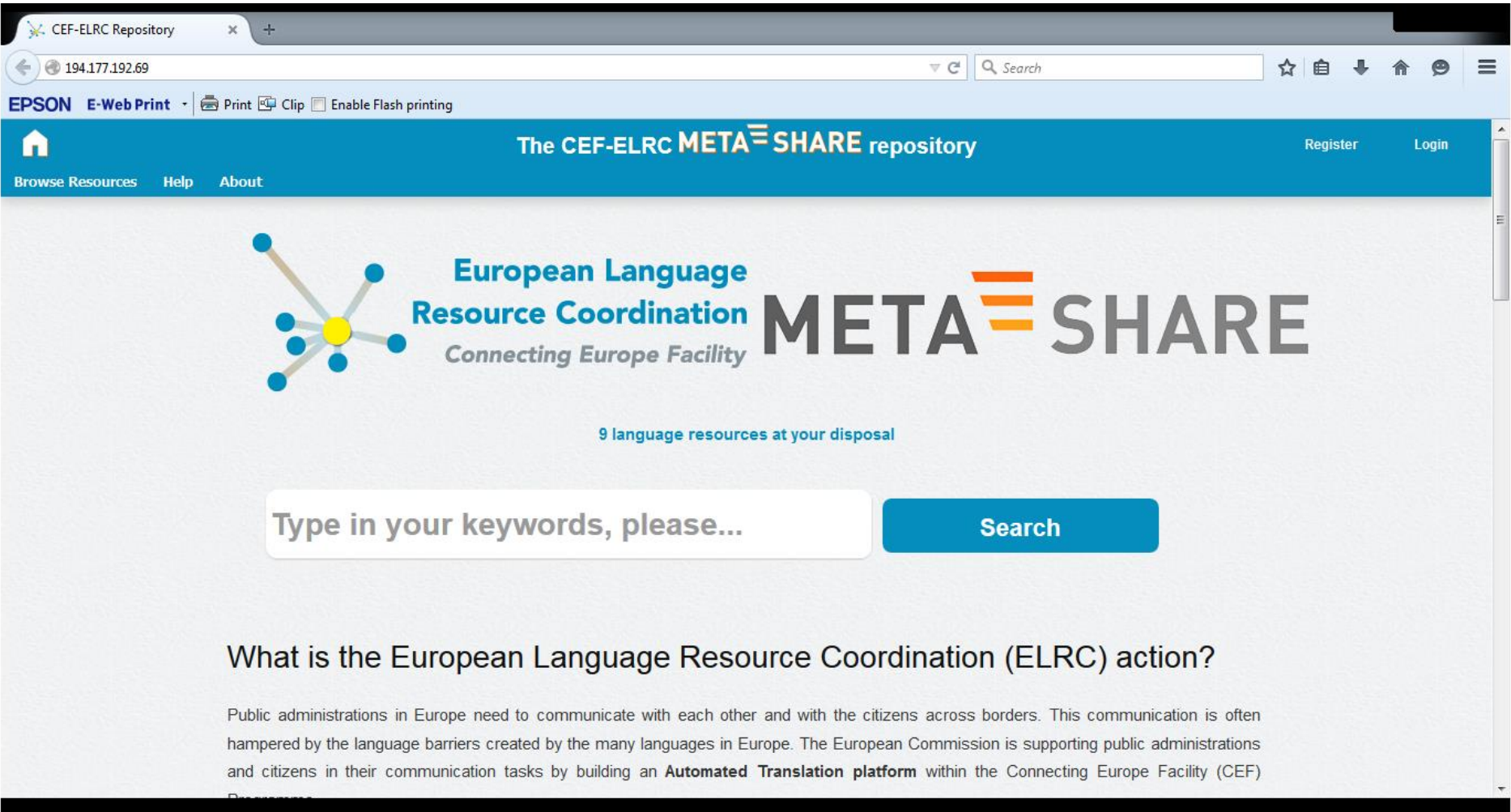


- [Home](#)
- [About](#)
- [News](#)
- [Helpdesk](#)
- [Events](#)
- [Resources](#)
- [Anchor Points](#)
- [Multilingual Europe](#)

European Language
Resource Coordination



Languages — the heart of
Multilingual Europe



The screenshot shows a web browser displaying the CEF-ELRC Repository website. The browser's address bar shows the URL 194.177.192.69. The website's header features the text "The CEF-ELRC META SHARE repository" and navigation links for "Register" and "Login". Below the header, there is a search bar with the placeholder text "Type in your keywords, please..." and a blue "Search" button. The main content area displays the European Language Resource Coordination logo and the text "9 language resources at your disposal". Below this, there is a section titled "What is the European Language Resource Coordination (ELRC) action?" followed by a paragraph of text.

CEP-ELRC Repository

194.177.192.69

EPSON E-Web Print Print Clip Enable Flash printing

The CEF-ELRC META SHARE repository

Register Login

Browse Resources Help About

European Language Resource Coordination Connecting Europe Facility

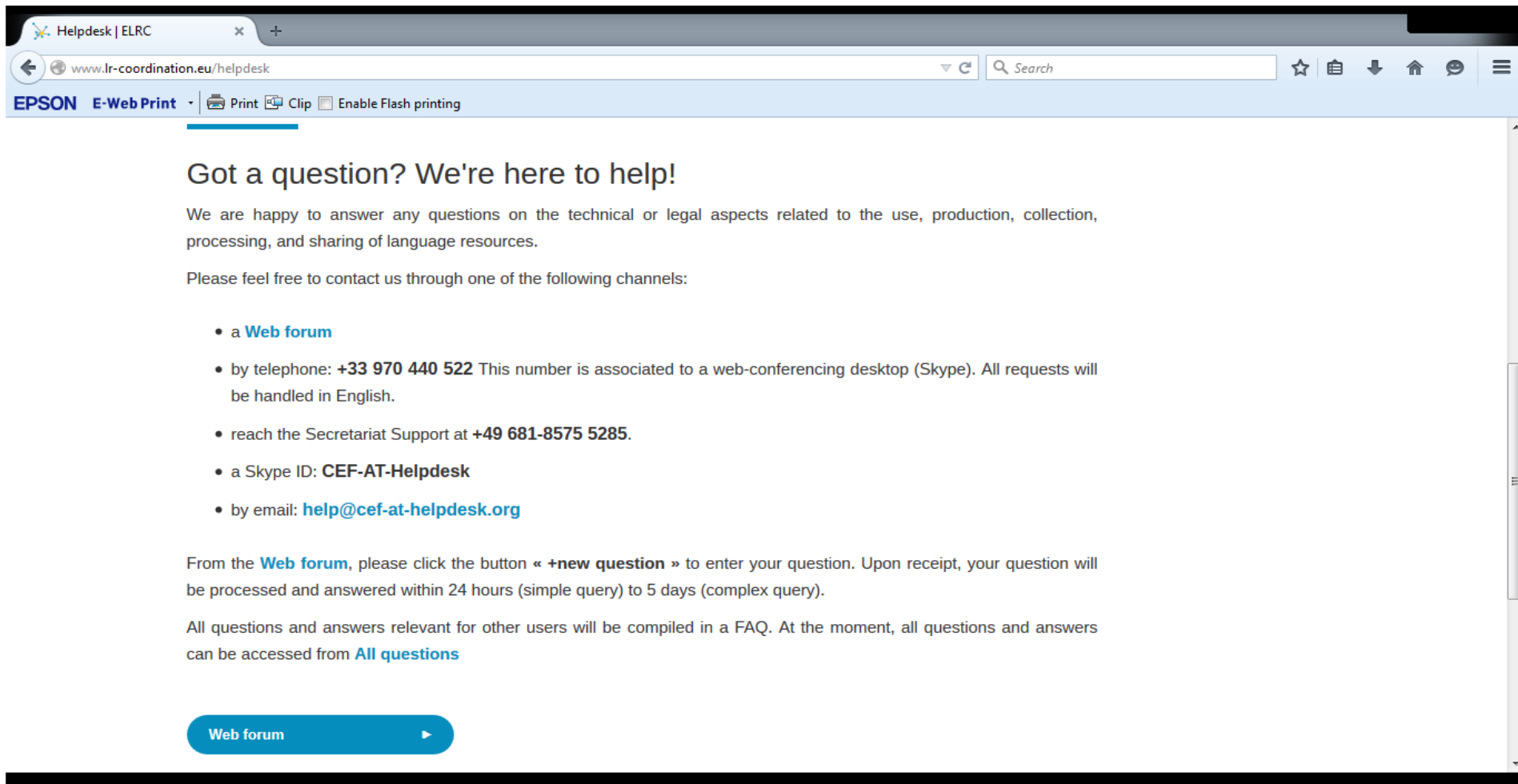
META SHARE

9 language resources at your disposal

Type in your keywords, please... Search

What is the European Language Resource Coordination (ELRC) action?

Public administrations in Europe need to communicate with each other and with the citizens across borders. This communication is often hampered by the language barriers created by the many languages in Europe. The European Commission is supporting public administrations and citizens in their communication tasks by building an **Automated Translation platform** within the Connecting Europe Facility (CEF)



The screenshot shows a web browser window with the URL www.lr-coordination.eu/helpdesk. The page features a navigation bar with the EPSON logo and 'E-Web Print' options. The main content area has a heading 'Got a question? We're here to help!' followed by a paragraph stating the helpdesk's purpose. Below this, a list of contact channels is provided, including a web forum, telephone numbers for Skype and Secretariat Support, a Skype ID, and an email address. A 'Web forum' button is located at the bottom of the content area.

Helpdesk | ELRC

www.lr-coordination.eu/helpdesk

EPSON E-Web Print Print Clip Enable Flash printing

Got a question? We're here to help!

We are happy to answer any questions on the technical or legal aspects related to the use, production, collection, processing, and sharing of language resources.

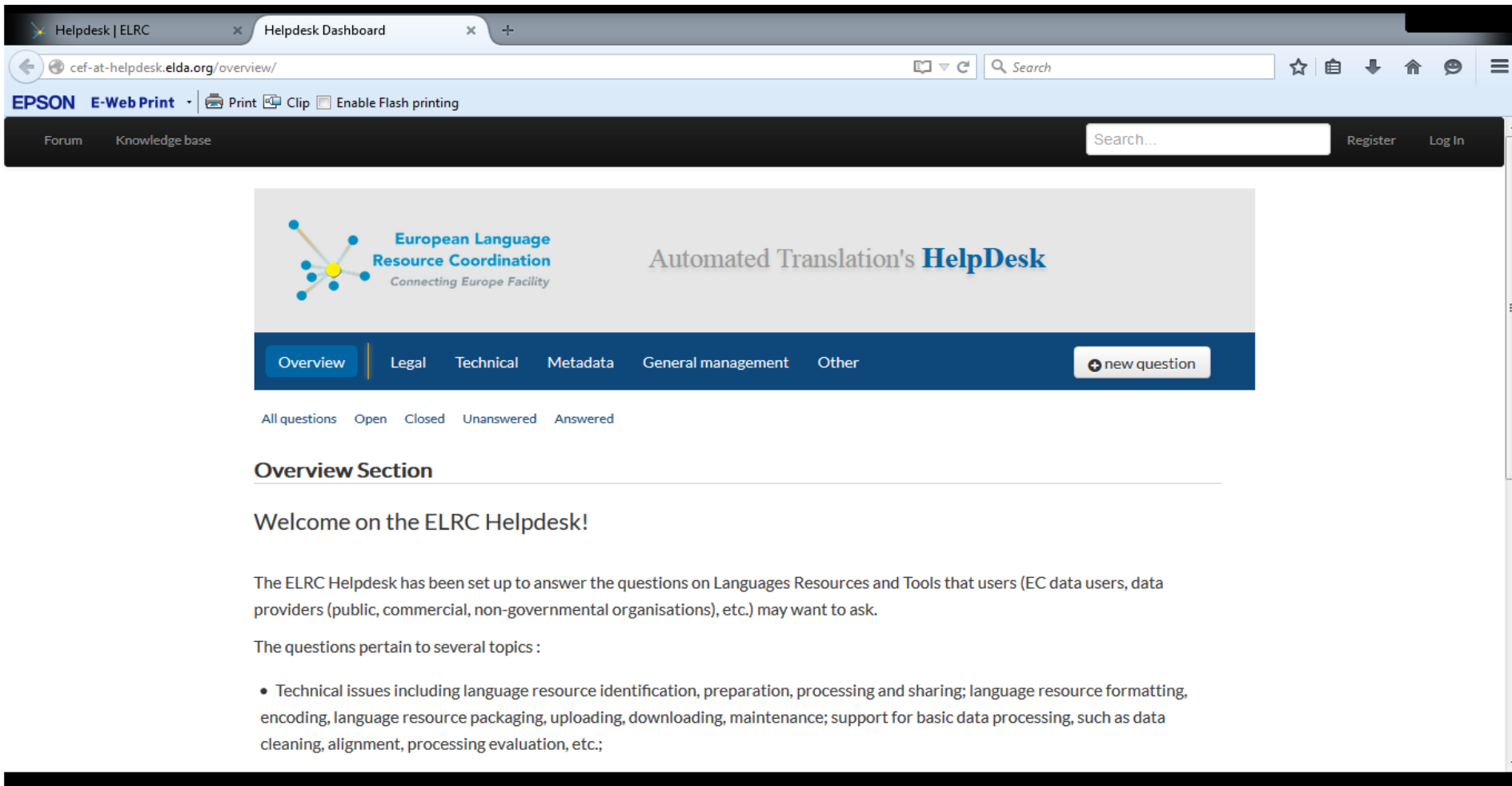
Please feel free to contact us through one of the following channels:

- a [Web forum](#)
- by telephone: **+33 970 440 522** This number is associated to a web-conferencing desktop (Skype). All requests will be handled in English.
- reach the Secretariat Support at **+49 681-8575 5285**.
- a Skype ID: **CEF-AT-Helpdesk**
- by email: help@cef-at-helpdesk.org

From the [Web forum](#), please click the button « **+new question** » to enter your question. Upon receipt, your question will be processed and answered within 24 hours (simple query) to 5 days (complex query).

All questions and answers relevant for other users will be compiled in a FAQ. At the moment, all questions and answers can be accessed from [All questions](#)

[Web forum](#)



The screenshot shows a web browser window with two tabs: 'Helpdesk | ELRC' and 'Helpdesk Dashboard'. The address bar shows 'cef-at-helpdesk.elda.org/overview/'. The page header includes 'EPSON E-Web Print', 'Print', 'Clip', and 'Enable Flash printing' options. A navigation bar contains 'Forum', 'Knowledge base', a search box, and 'Register' and 'Log In' links. The main content area features the ELRC logo and the title 'Automated Translation's HelpDesk'. Below this is a navigation menu with 'Overview', 'Legal', 'Technical', 'Metadata', 'General management', and 'Other' tabs, along with a '+ new question' button. A filter bar shows 'All questions', 'Open', 'Closed', 'Unanswered', and 'Answered' options. The 'Overview Section' is titled 'Welcome on the ELRC Helpdesk!' and contains a paragraph explaining the helpdesk's purpose and a list of topics it covers.

Helpdesk | ELRC x Helpdesk Dashboard x +

cef-at-helpdesk.elda.org/overview/ Search

EPSON E-Web Print Print Clip Enable Flash printing

Forum Knowledge base Search... Register Log In

European Language Resource Coordination Connecting Europe Facility

Automated Translation's HelpDesk

Overview Legal Technical Metadata General management Other + new question

All questions Open Closed Unanswered Answered

Overview Section

Welcome on the ELRC Helpdesk!

The ELRC Helpdesk has been set up to answer the questions on Languages Resources and Tools that users (EC data users, data providers (public, commercial, non-governmental organisations), etc.) may want to ask.

The questions pertain to several topics :

- Technical issues including language resource identification, preparation, processing and sharing; language resource formatting, encoding, language resource packaging, uploading, downloading, maintenance; support for basic data processing, such as data cleaning, alignment, processing evaluation, etc.;