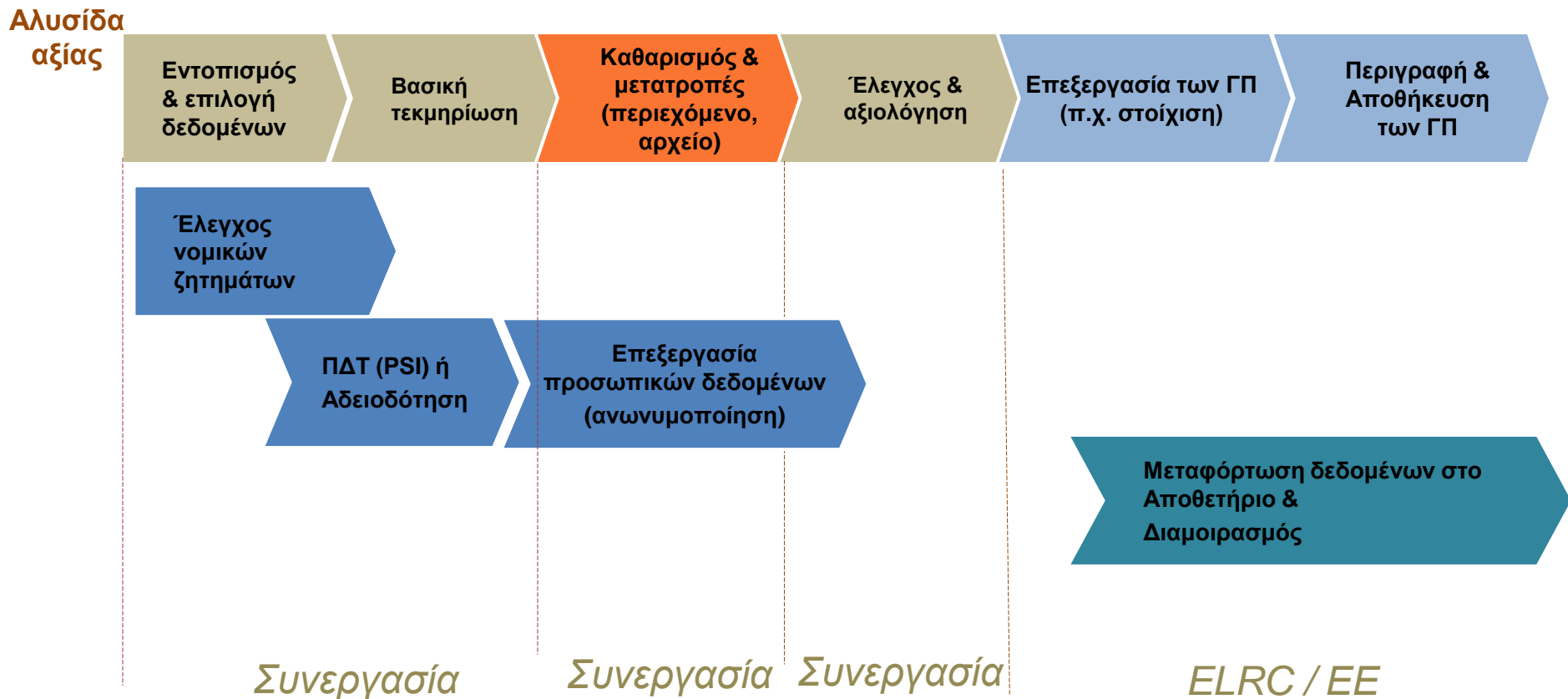


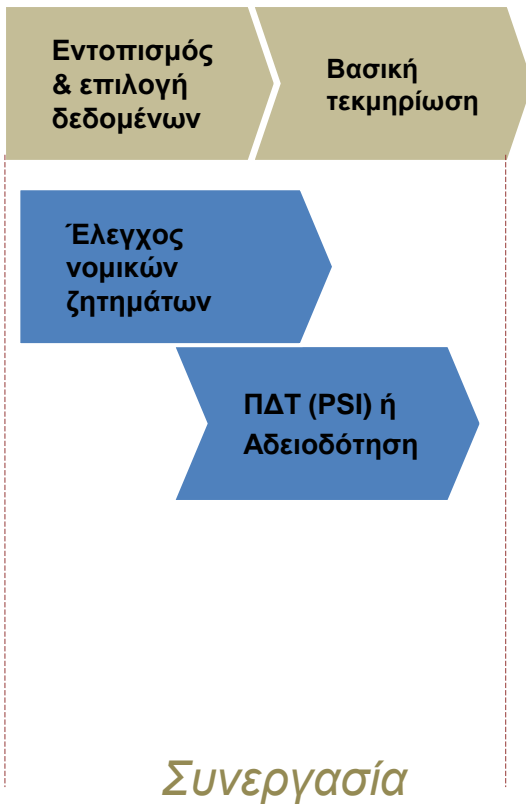
# Διαμοιρασμός δεδομένων & γλωσσικών πόρων: τεχνικά ζητήματα

Πένυ Λαμπροπούλου  
ΙΕΛ/ΕΚ "ΑΘΗΝΑ"

- Γνωρίζετε ήδη “*Τι είδους δεδομένα χρειάζονται*” από προηγούμενη συνεδρία
- Μοντέλο που βασίζεται στα Δεδομένα (Data Driven Paradigm)
- Χρειαζόμαστε δεδομένα στην ελληνική αλλά και σε όλες τις γλώσσες
- Πού μπορούμε να εντοπίσουμε δεδομένα:
  - σε Φορείς του Δημοσίου
    - Ορατά Δεδομένα, π.χ. διαδίκτυο (σελίδες HTML, αναφορές, κτλ.)
    - Κρυμμένα/Αόρατα Δεδομένα: αρχεία/συλλογές, κρυμμένο διαδίκτυο (deep web), εσωτερικά αποθετήρια φορέων
    - Μέσω των Παρόχων Γλωσσικών Υπηρεσιών με τους οποίους συνεργάζεστε

# Ροή εργασίας για τη συλλογή & επεξεργασία των δεδομένων Δεδομένα → Γλωσσικοί Πόροι (ΓΠ)





- Εντοπισμός πηγών
- Εντοπισμός και επιλογή **συνόλων δεδομένων** (μη επεξεργασμένων)
- Έλεγχος νομικών θεμάτων
  - [Άδεια χρήσης](#)
  - Διαχείριση προσωπικών & ευαίσθητων δεδομένων
- Έλεγχος τεχνικών θεμάτων
  - [Επιλογή μορφοτύπων](#) των "μη επεξεργασμένων" δεδομένων
- Τεκμηρίωση των δεδομένων με την ελάχιστη βασική πληροφορία (Τίτλος, Γλώσσες, Θεματικά πεδία, ...)



- Θέματα που επηρεάζουν τη διαδικασία συλλογής δεδομένων
  - εξ ορισμού ανοιχτά δεδομένα, π.χ. ΠΔΤ
  - δεδομένα για τη χρήση των οποίων απαιτείται αίτηση
- Θέματα αδειοδότησης
  - μπορούμε να σας βοηθήσουμε στις διαδικασίες
  - πρότυπες άδειες χρήσης
    - Κυβερνητικές Ανοιχτές Άδειες (Government Open Licenses)
    - Πρότυπες Άδειες Επαναχρησιμοποίησης
    - Διαλειτουργικότητα αδειών



# Δεδομένα σε ψηφιακή μορφή και σε κάθε είδους μορφότυπο!!



**Προτιμητέα μορφότυπα:**  
αυτά που μπορούν να  
χρησιμοποιηθούν από τα  
εργαλεία επεξεργασίας και  
την πλατφόρμα CEF.AT)





Καθαρισμός &  
Μετατροπές  
(περιεχόμενο,  
αρχείο)

Επεξεργασία  
προσωπικών δεδομένων  
(ανωνυμοποίηση)

Συνεργασία

### Τεχνικά θέματα

- "Καθαρισμός" των αρχείων
  - μετατροπή χαρακτήρων σε κωδικοποίηση UTF8
  - [αφαίρεση μορφοποίησης](#) (π.χ. εντολών για έντονα γράμματα / πλαγιογράμματα, γραφικών, πινάκων, διαφημίσεων, ετικετών html κτλ.)
  - ...
- Μετατροπή των αρχείων στο απαιτούμενο / κατάλληλο μορφότυπο (π.χ. μετατροπή σε XML, XLIFF κτλ.)
- [Ανωνυμοποίηση δεδομένων](#) (αν και όπου χρειάζεται και, βεβαίως, αν είναι εφικτό)





***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 πρώτους στην





- Εντοπίζουμε μια μεγάλη πηγή δεδομένων με στοιχεία για άτομα, φορείς κτλ.
- Μπορούμε πάντοτε να χρησιμοποιήσουμε εργαλεία Αναγνώρισης Ονομάτων (Named Entity Recognizer) για να βρούμε τα (προσωπικά) στοιχεία (ονόματα, τοποθεσίες, ημερομηνίες, πληροφορίες γέννησης, κτλ.), να τα αφαιρέσουμε και να τα αντικαταστήσουμε με σύμβολα
- Επιβεβαιώνουμε ότι τα αποτελέσματα της επεξεργασίας καλύπτουν τις απαιτήσεις του παρόχου – εάν η ανωνυμοποίηση δεν έχει τα απαιτούμενα ποσοστά επιτυχίας, απορρίπτουμε το συγκεκριμένο σύνολο δεδομένων





Έλεγχος &  
Αξιολόγηση

- Έλεγχος αποτελεσμάτων ανωνυμοποίησης
- Αξιολόγηση & Έλεγχος ποιότητας των δεδομένων μετά την επεξεργασία (μορφότυπο & περιεχόμενο ΓΠ)

→ αποδοχή / απόρριψη ΓΠ

*Συνεργασία*



Επεξεργασία  
των ΓΠ  
(π.χ. στοίχιση)

Περιγραφή &  
Αποθήκευση  
των ΓΠ

- Προετοιμασία και επεξεργασία των ΓΠ για τα εργαλεία Αυτόματης Μετάφρασης (π.χ. στοίχιση)
- Τεκμηρίωση ΓΠ (προσθήκη μεταδεδομένων για την περιγραφή του ΓΠ)
- Μεταφόρτωση του ΓΠ σε ειδικό [Αποθετήριο Δεδομένων](#) που υποστηρίζει τον διαμοιρασμό τους

Μεταφόρτωση Δεδομένων στο  
Αποθετήριο &  
Διαμοιρασμός

*ELRC / EE*

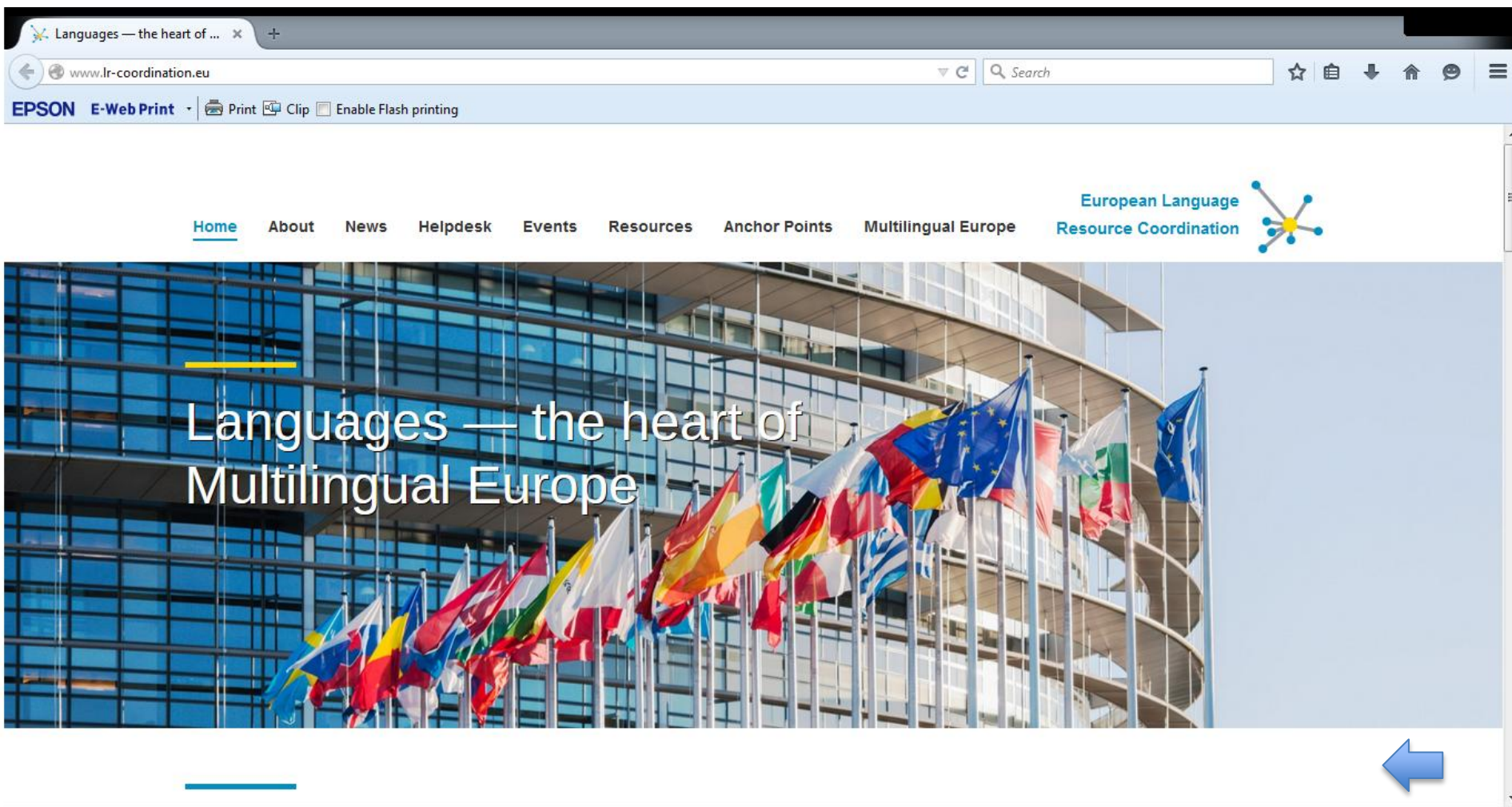
- Γνωρίζετε τα δεδομένα σας
  - και τα ορατά και τα αόρατα
- Η πρόσβαση στα αρχεία, στο "κρυμμένο" διαδίκτυο δεν είναι δυνατή για τους εκτός του φορέα
- Δεν ισχύει για όλα τα δεδομένα η οδηγία ΠΔΤ ή κάποια επιτρεπτική άδεια χρήσης
- Η πρόσβαση σε παράγωγες μορφές (π.χ. PDF) είναι λιγότερο αποδοτική από την πρόσβαση στο περιεχόμενο στην αρχική του μορφή (το οποίο μπορεί να βρίσκεται αποθηκευμένο στα αποθετήρια του φορέα και να έχουν πρόσβαση σε αυτό μόνο οι εργαζόμενοι του φορέα)

# Δράσεις σε συνεργασία με τους φορείς

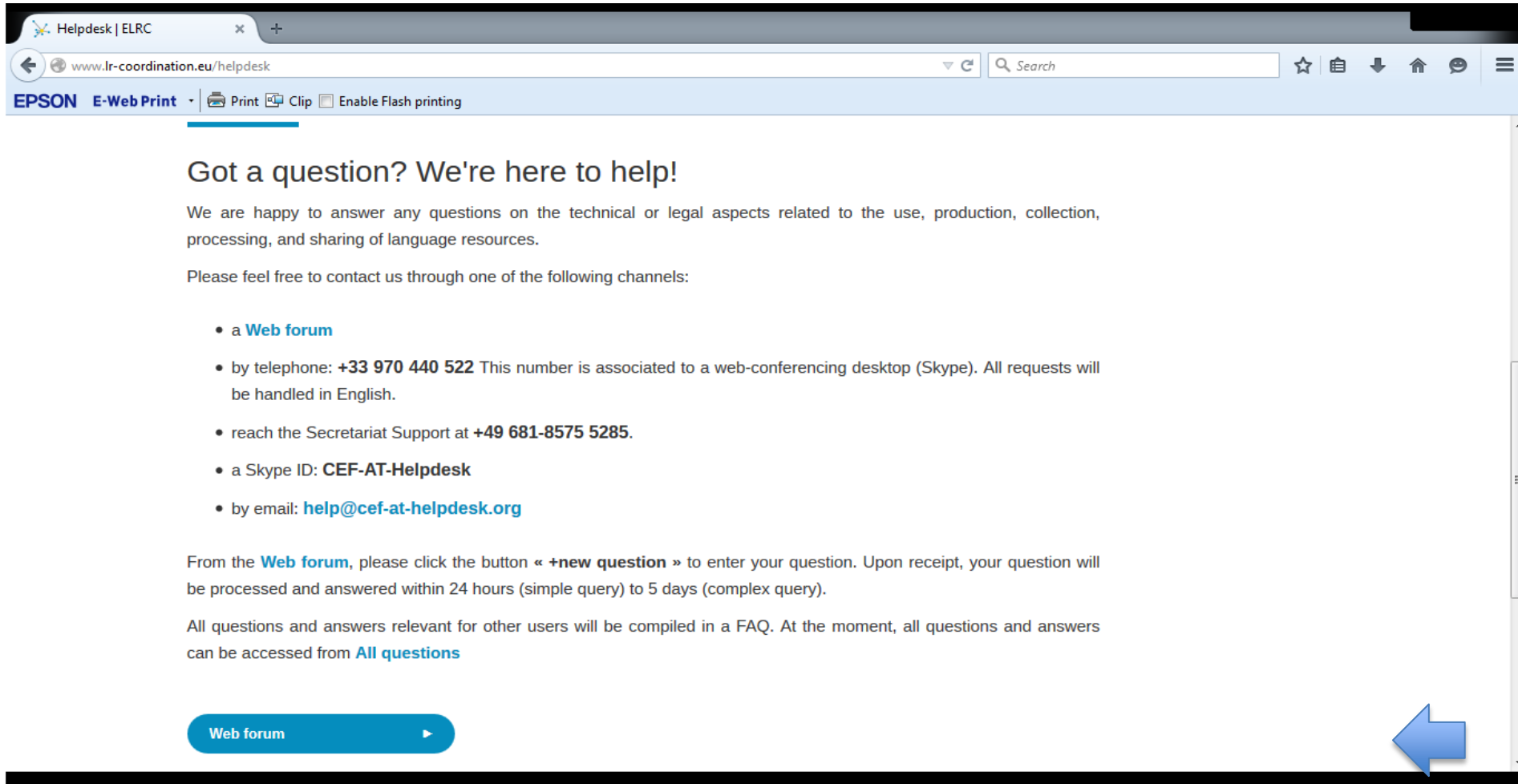


- Εντοπισμός & επιλογή πηγών
- Εντοπισμός & επιλογή συνόλων δεδομένων (μη επεξεργασμένων)
  - Τα δεδομένα μπορούν να αντληθούν από ορατές πηγές (π.χ. να συλλεχθούν από το διαδίκτυο με αυτόματες μεθόδους)
  - Τα δεδομένα μπορούν να τα προσφέρουν οι ίδιοι οι φορείς του δημοσίου
  - Οι φορείς του δημοσίου μπορούν να ενισχύσουν τον εντοπισμό των ορατών πηγών
- Οι δράσεις επεξεργασίας που παρουσιάστηκαν μπορούν να γίνουν από κοινού από την κοινοπραξία ELRC και τον πάροχο δεδομένων

- Υποστήριξη σε όλες τις διαδικασίες και σε πρακτικά θέματα με τις ακόλουθες υπηρεσίες
  - [ιστότοπος ELRC](#)
  - [γραφεία τεχνικής & νομικής υποστήριξης](#)
  - [αποθετήριο διαμοιρασμού γλωσσικών πόρων](#)
  - [φόρουμ χρηστών](#)



The screenshot shows a web browser window displaying the ELRC website. The browser's address bar shows the URL [www.lr-coordination.eu](http://www.lr-coordination.eu). The website's navigation menu includes links for Home, About, News, Helpdesk, Events, Resources, Anchor Points, and Multilingual Europe. The main content area features a large banner image of various European national flags and the European Union flag in front of a modern building. Overlaid on this image is the text "Languages — the heart of Multilingual Europe". The website logo is visible in the top right corner of the page content.



The screenshot shows a web browser window with the URL [www.lr-coordination.eu/helpdesk](http://www.lr-coordination.eu/helpdesk). The page content includes:

## 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:

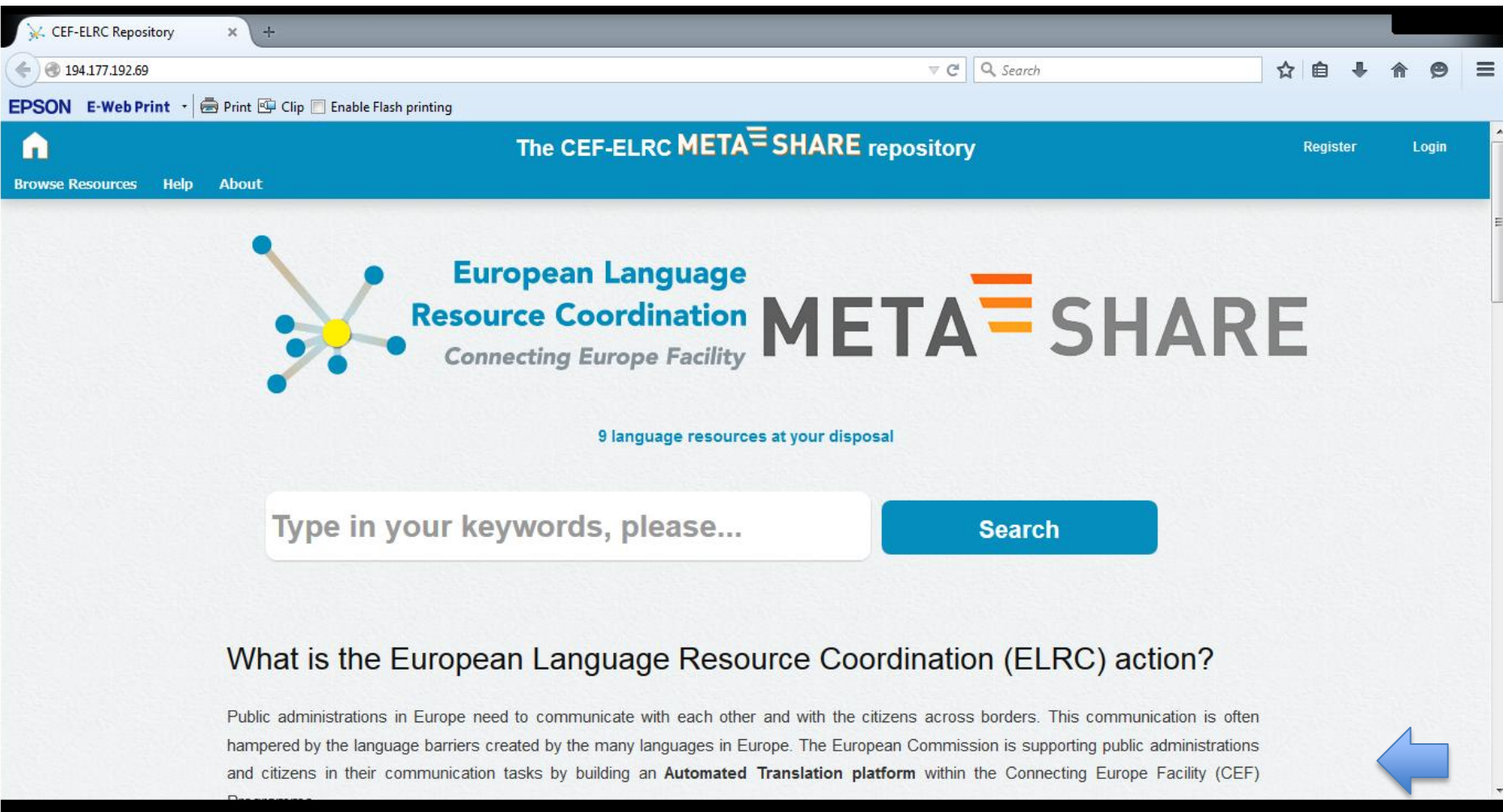
- a [Web forum](#)
- by telephone: **+33 970 440 522** This number is associated to a web-conferencing desktop (Skype). All requests will be handled in English.
- reach the Secretariat Support at **+49 681-8575 5285**.
- a Skype ID: **CEF-AT-Helpdesk**
- by email: [help@cef-at-helpdesk.org](mailto:help@cef-at-helpdesk.org)

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](#)

At the bottom of the page, there is a blue button labeled "Web forum" with a right-pointing arrow, and a large blue arrow pointing left.





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 includes 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 features 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. A blue arrow points to the right in the bottom right corner of the screenshot.

CEF-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

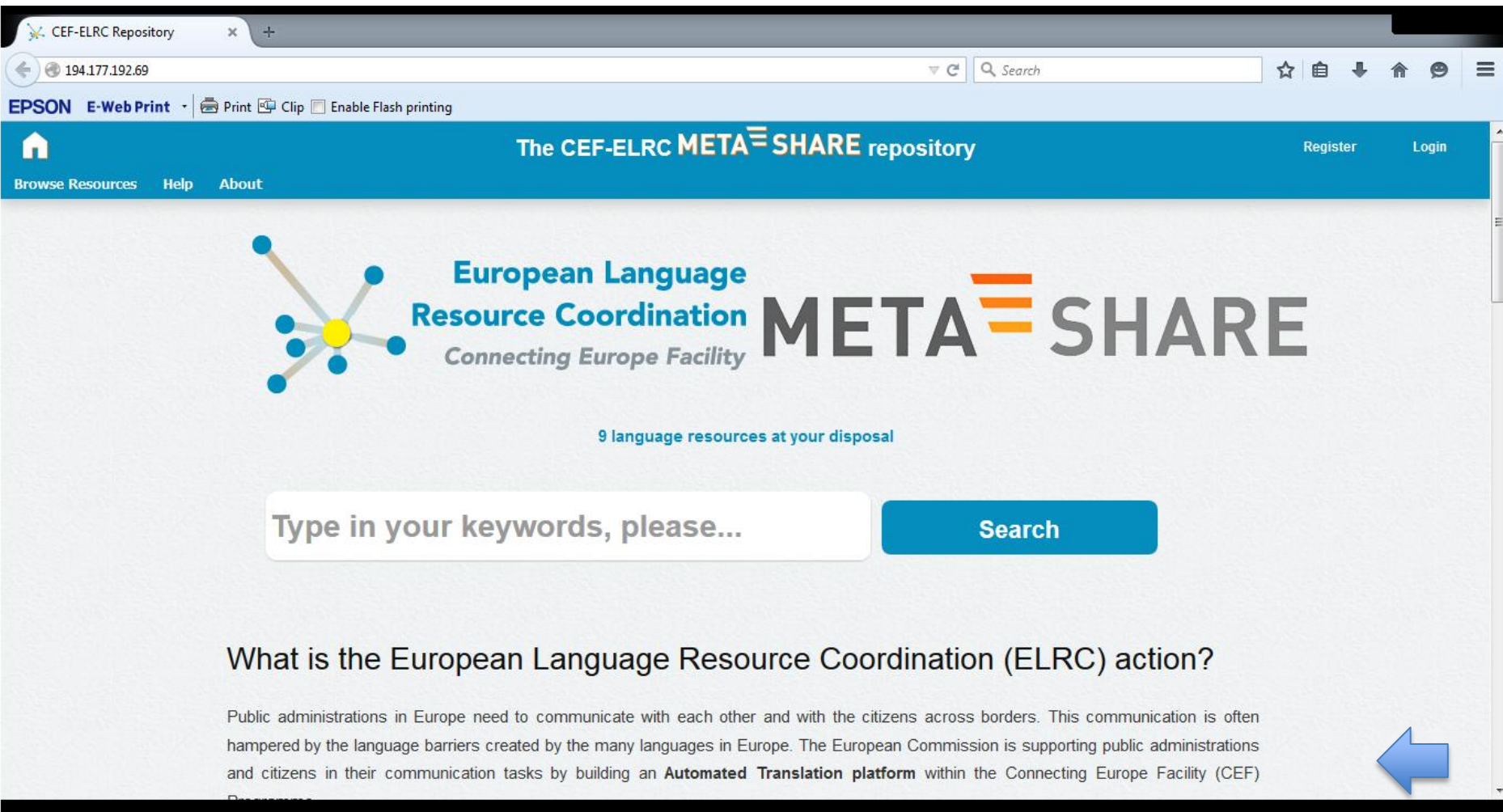
META SHARE

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 displaying the CEF-ELRC Repository website. The browser's address bar shows the URL 194.177.192.69. The website's header is blue and contains 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 features 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. A blue arrow points to the right in the bottom right corner of the screenshot.

CEF-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

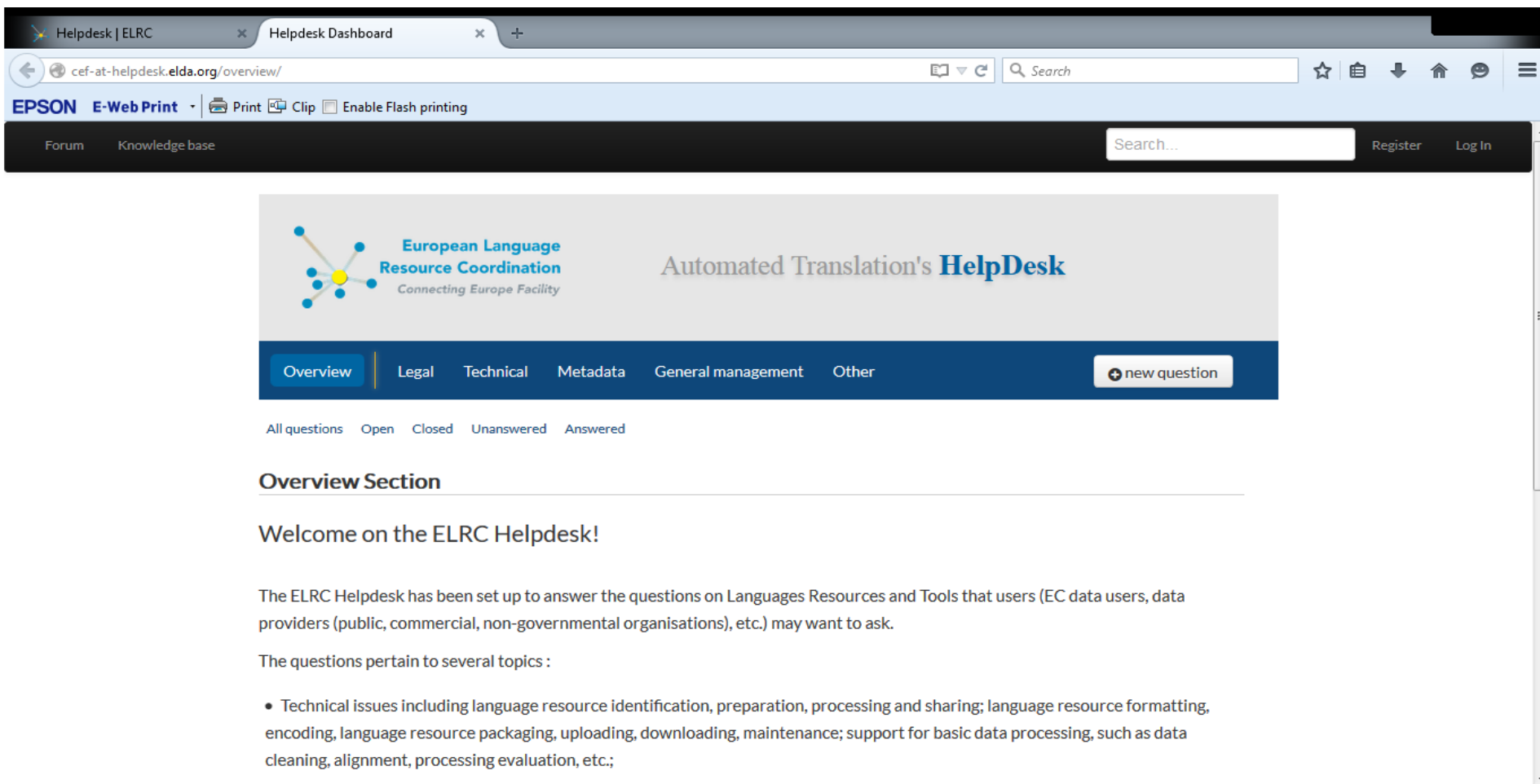
META SHARE

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 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, '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, and 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 welcome message and a list of topics.

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.;

- Η αξιοποίηση με τις κατάλληλες προσαρμογές υπαρχόντων δεδομένων (π.χ. μεταφράσεων που έχουν γίνει από ανθρώπους) είναι ο καλύτερος τρόπος βελτίωσης της ποιότητας της Αυτόματης Μετάφρασης
- Τα μοντέλα που βασίζονται σε Δεδομένα προσφέρουν τον βέλτιστο τρόπο για να αποκτήσουν αξία οι υπάρχοντες πόροι
- Το ELRC μπορεί να σας βοηθήσει να ελέγξετε την καταλληλότητα των δεδομένων σας (σε οποιαδήποτε φάση)
- Ας αναγνωρίσουμε και ας δώσουμε στους γλωσσικούς σας πόρους την αξία που πραγματικά έχουν - σχεδιάστε από τώρα ένα Σχέδιο Διαχείρισης Δεδομένων