European Language Resource Coordination
Connecting Europe Facility

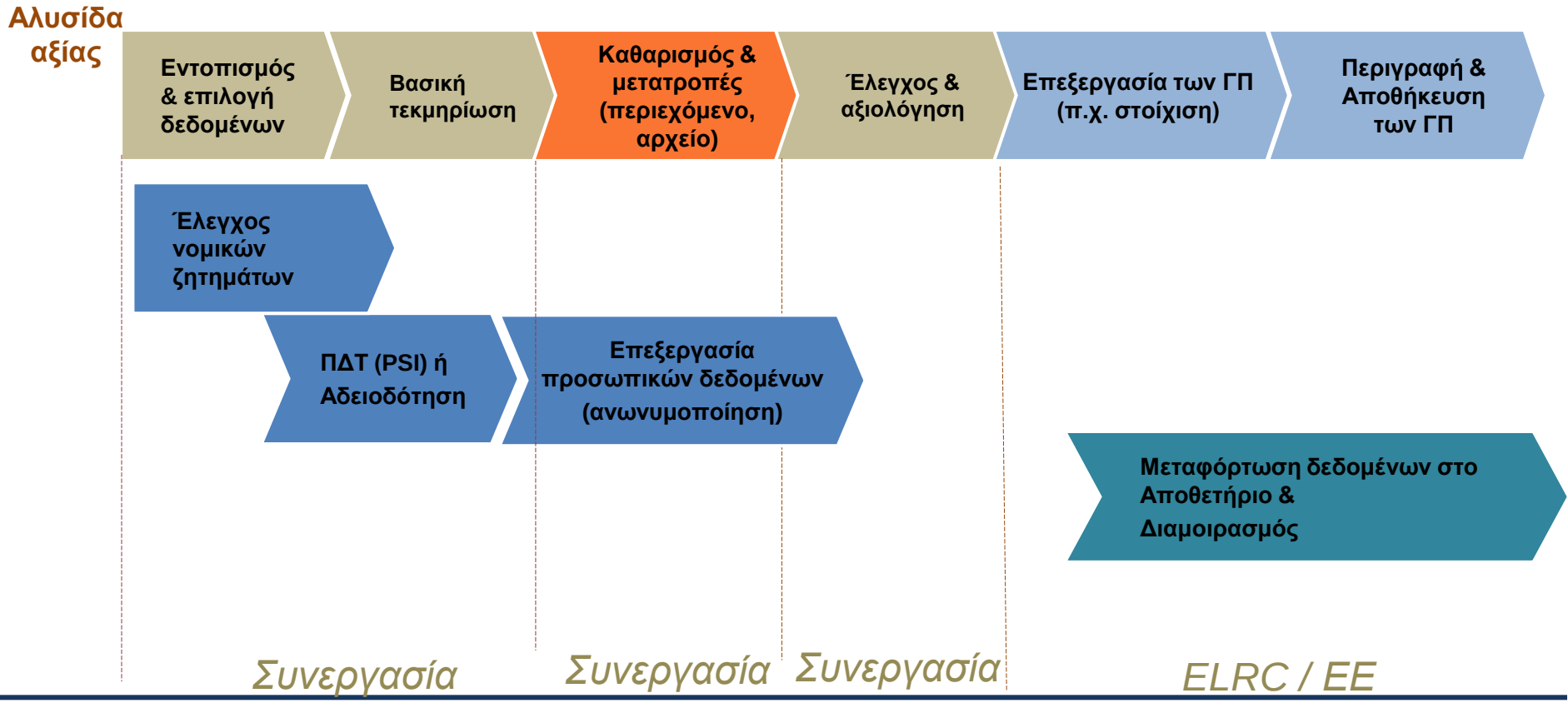
Διαμοιρασμός δεδομένων & γλωσσικών πόρων: τεχνικά ζητήματα

Πένυ Λαμπροπούλου
ΙΕΛ/ΕΚ "ΑΘΗΝΑ"

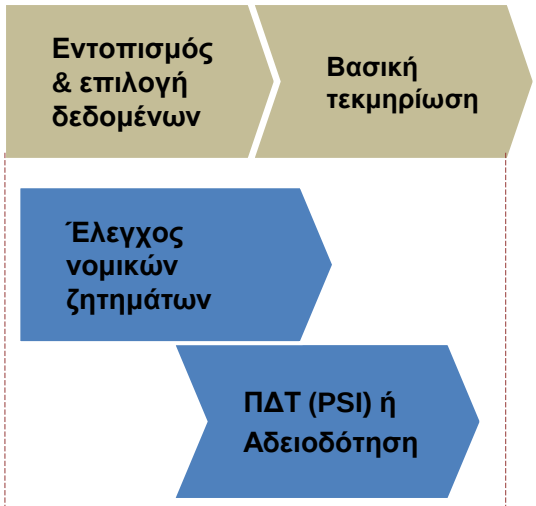
Επισκόπηση

- 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)



Θέματα προς εξέταση (1)



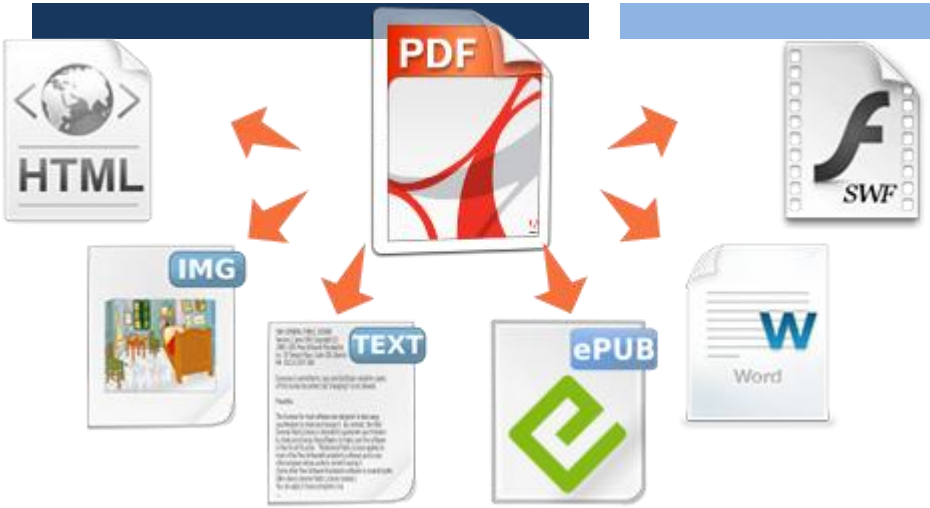
- Εντοπισμός πηγών
- Εντοπισμός και επιλογή **συνόλων δεδομένων** (μη επεξεργασμένων)
- Έλεγχος νομικών θεμάτων
  - Άδεια χρήσης
  - Διαχείριση προσωπικών & ευαίσθητων δεδομένων
- Έλεγχος τεχνικών θεμάτων
  - Επιλογή μορφοτύπων των "μη επεξεργασμένων" δεδομένων
- Τεκμηρίωση των δεδομένων με την ελάχιστη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

Νομικά θέματα

- 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)

Θέματα προς εξέταση (2)

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)

Παράδειγμα μορφοποίησης

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

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
ώπους των
οποίοι
σεις. Ο Διεθνής
ένα από τα
ποίος
στη

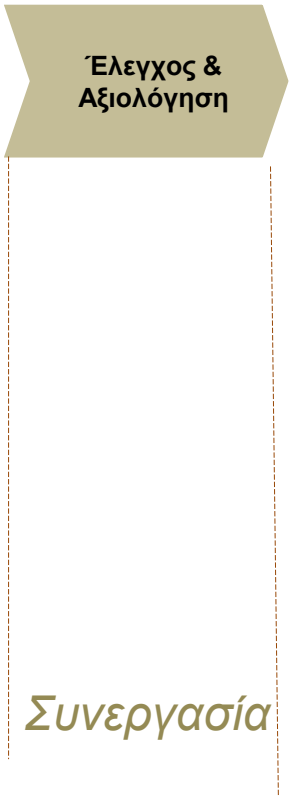
Greece is a place of culture, the arts and sciences. Its tradition of contribution to global cultural and scientific communities, combined with its outstanding natural beauty and excellent infrastructure, has made it an ideal place in which to hold conferences. Over the last few years, Greece has more 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 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.

Ανωνυμοποίηση δεδομένων

- 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

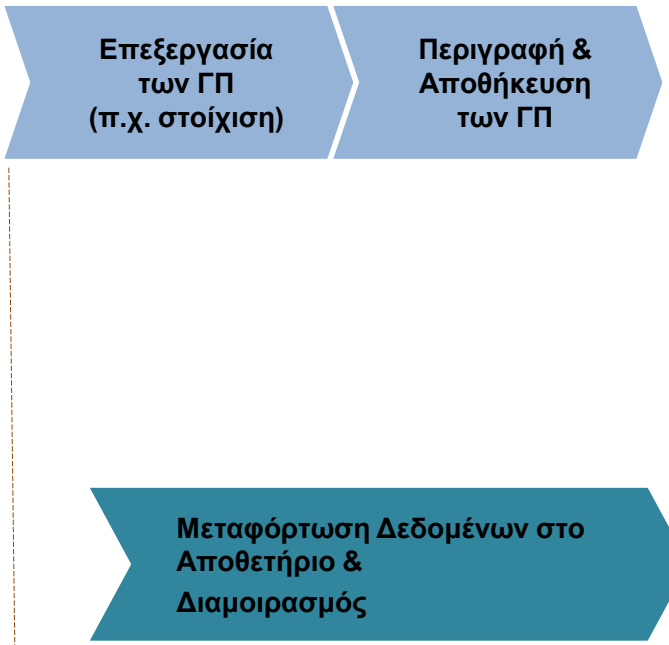
Θέματα προς εξέταση (3)



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

➔ αποδοχή / απόρριψη ΓΠ

Θέματα προς εξέταση (4)



- 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

European Language Resource Coordination
Connecting Europe Facility

Θέλουμε τη συμμετοχή σας

- 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

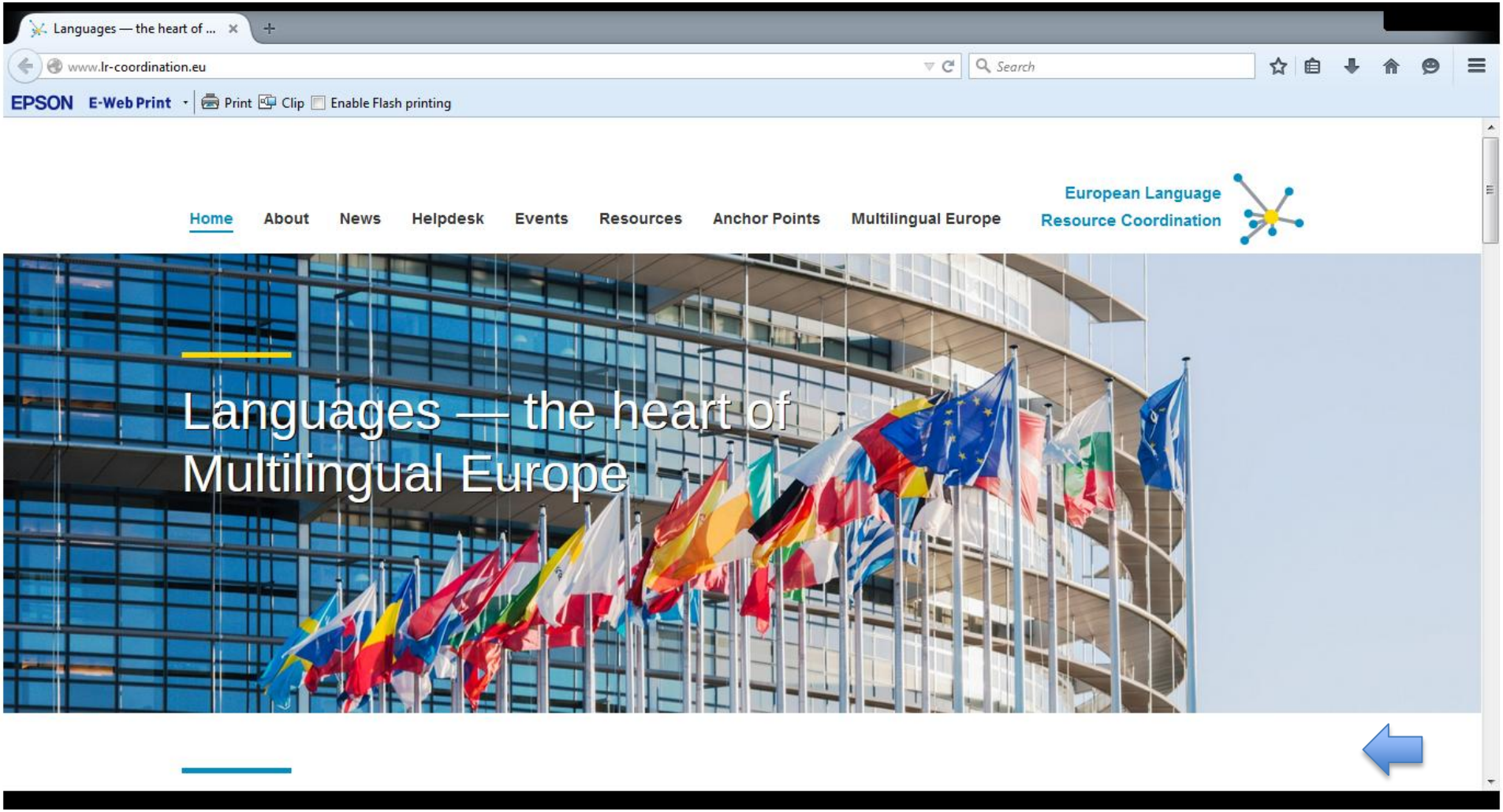
- Εντοπισμός & επιλογή πηγών
- Εντοπισμός & επιλογή συνόλων δεδομένων (μη επεξεργασμένων)
  - Τα δεδομένα μπορούν να αντληθούν από ορατές πηγές (π.χ. να συλλεχθούν από το διαδίκτυο με αυτόματες μεθόδους)
  - Τα δεδομένα μπορούν να τα προσφέρουν οι ίδιοι οι φορείς του δημοσίου
  - Οι φορείς του δημοσίου 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

Ρόλος της κοινοπραξίας ELRC

- 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

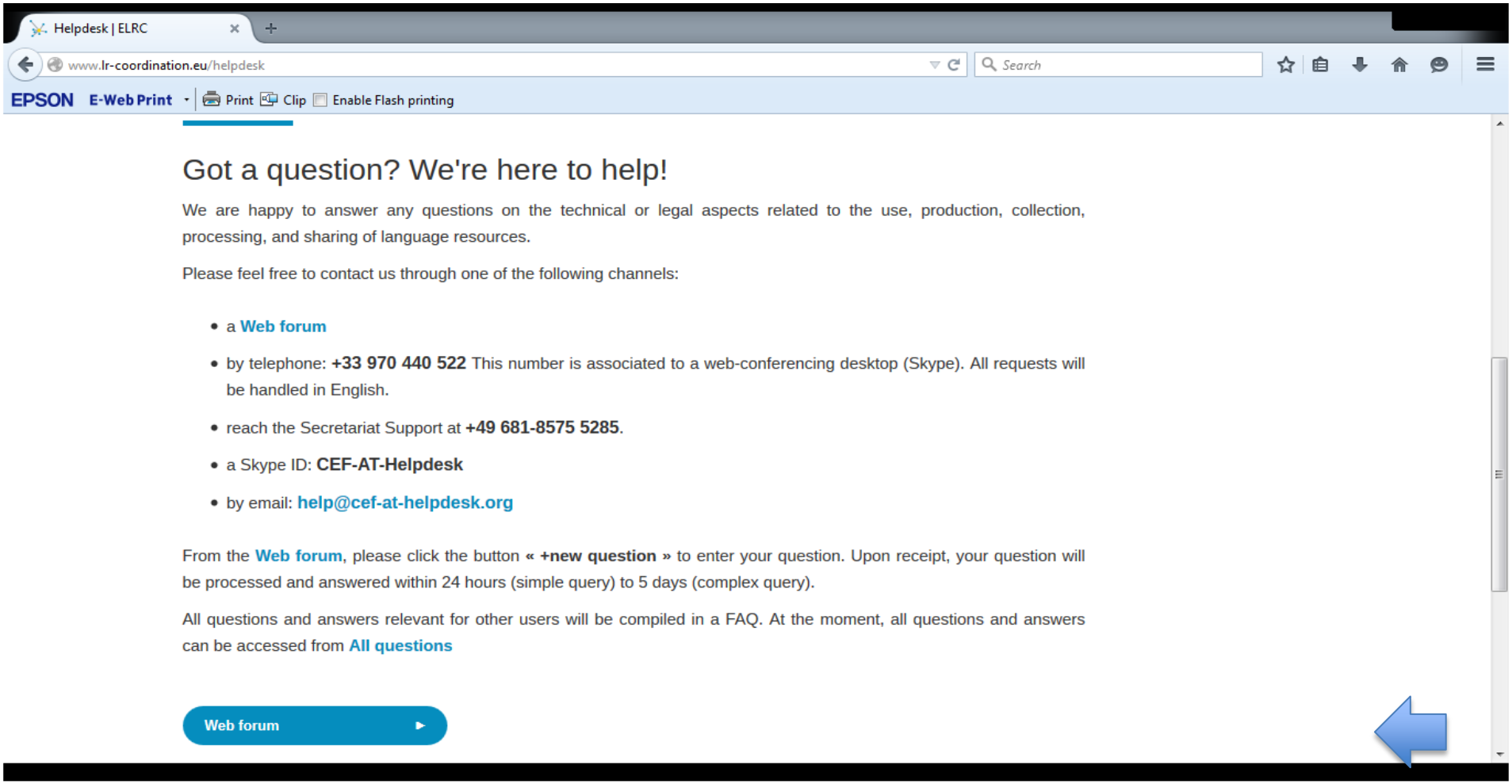
European Language Resource Coordination
Connecting Europe Facility

Ιστότοπος ELRC

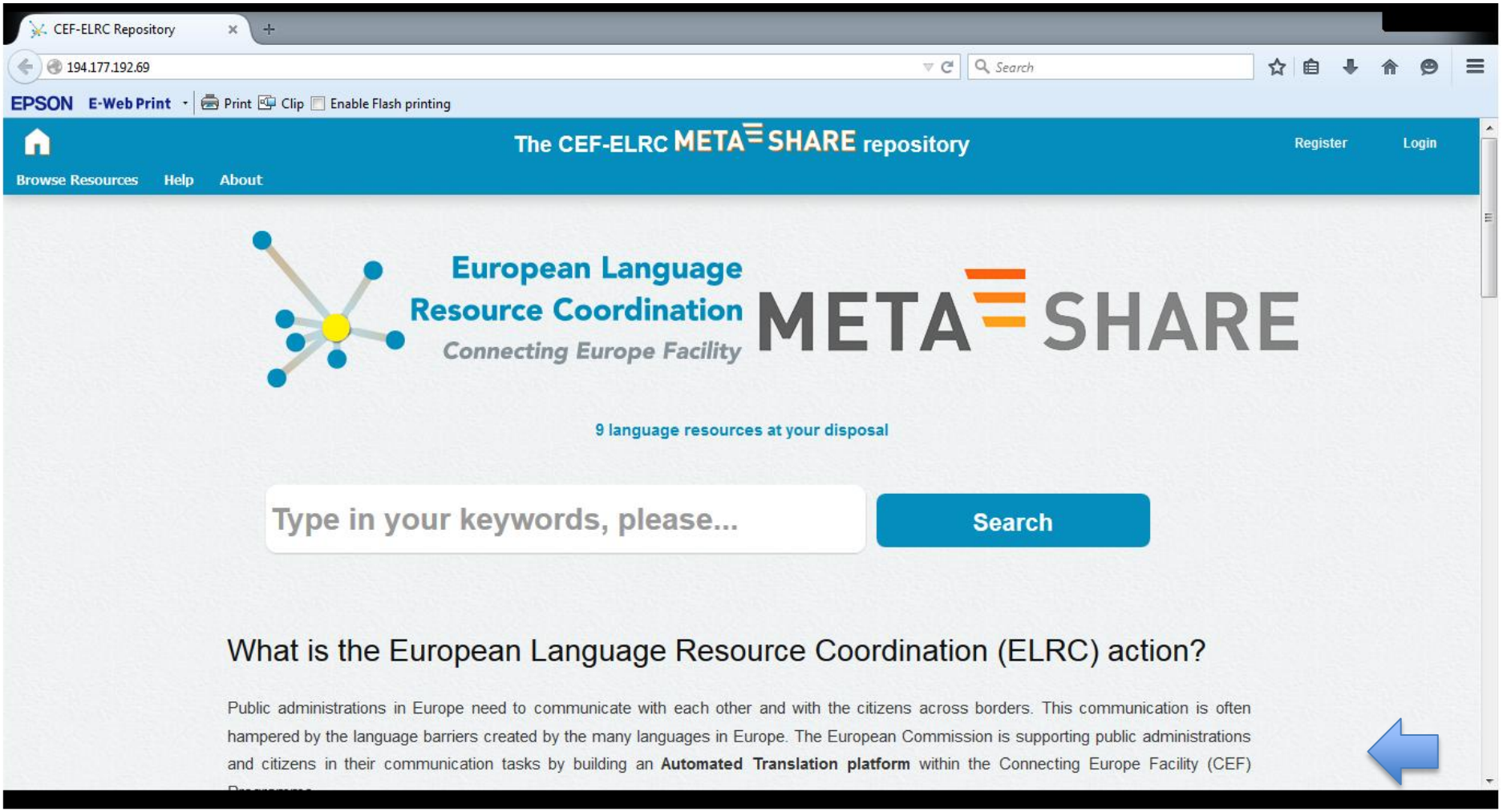


European Language Resource Coordination
Connecting Europe Facility

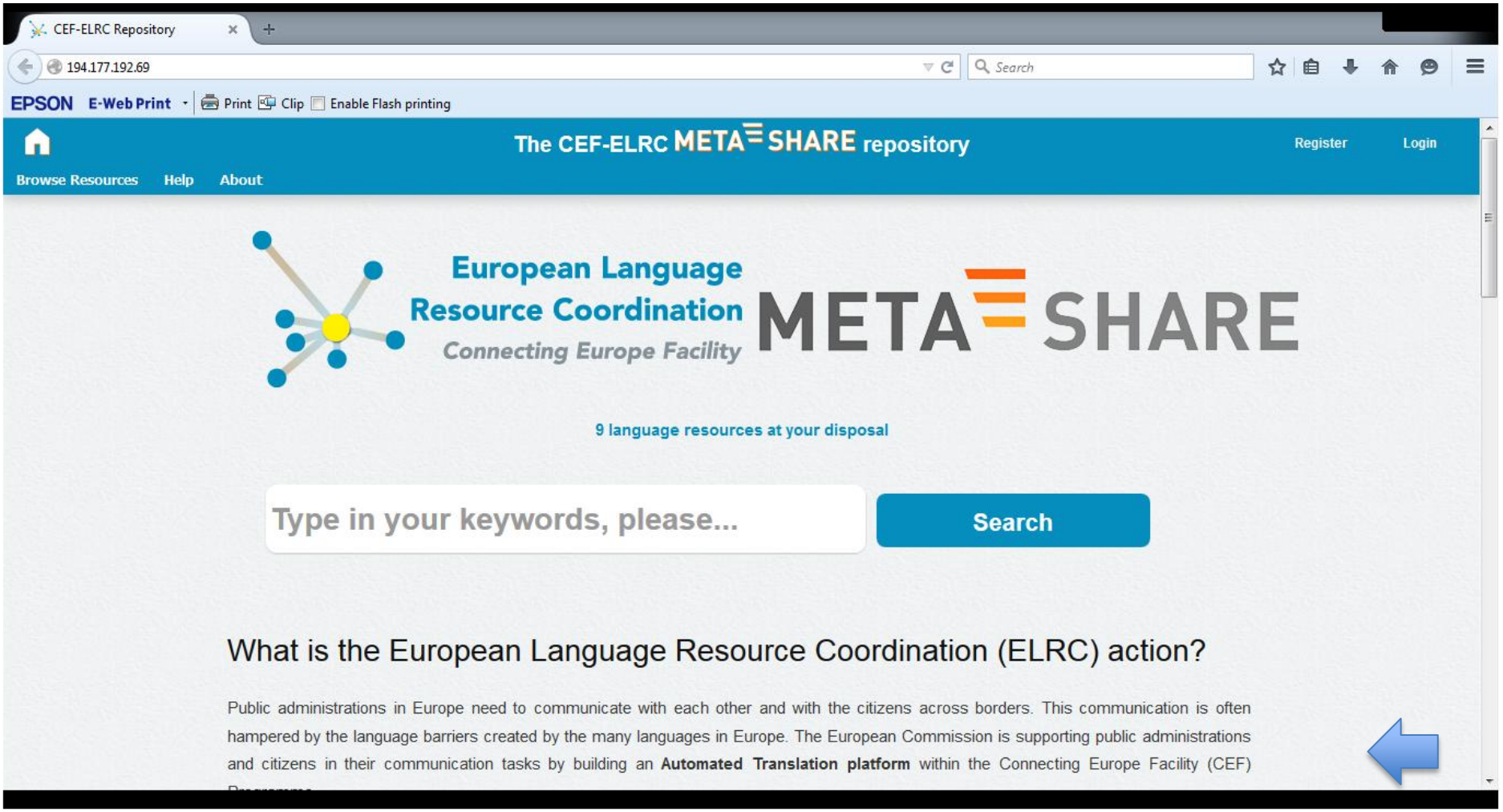
Ιστότοπος ELRC: Γραφείο υποστήριξης



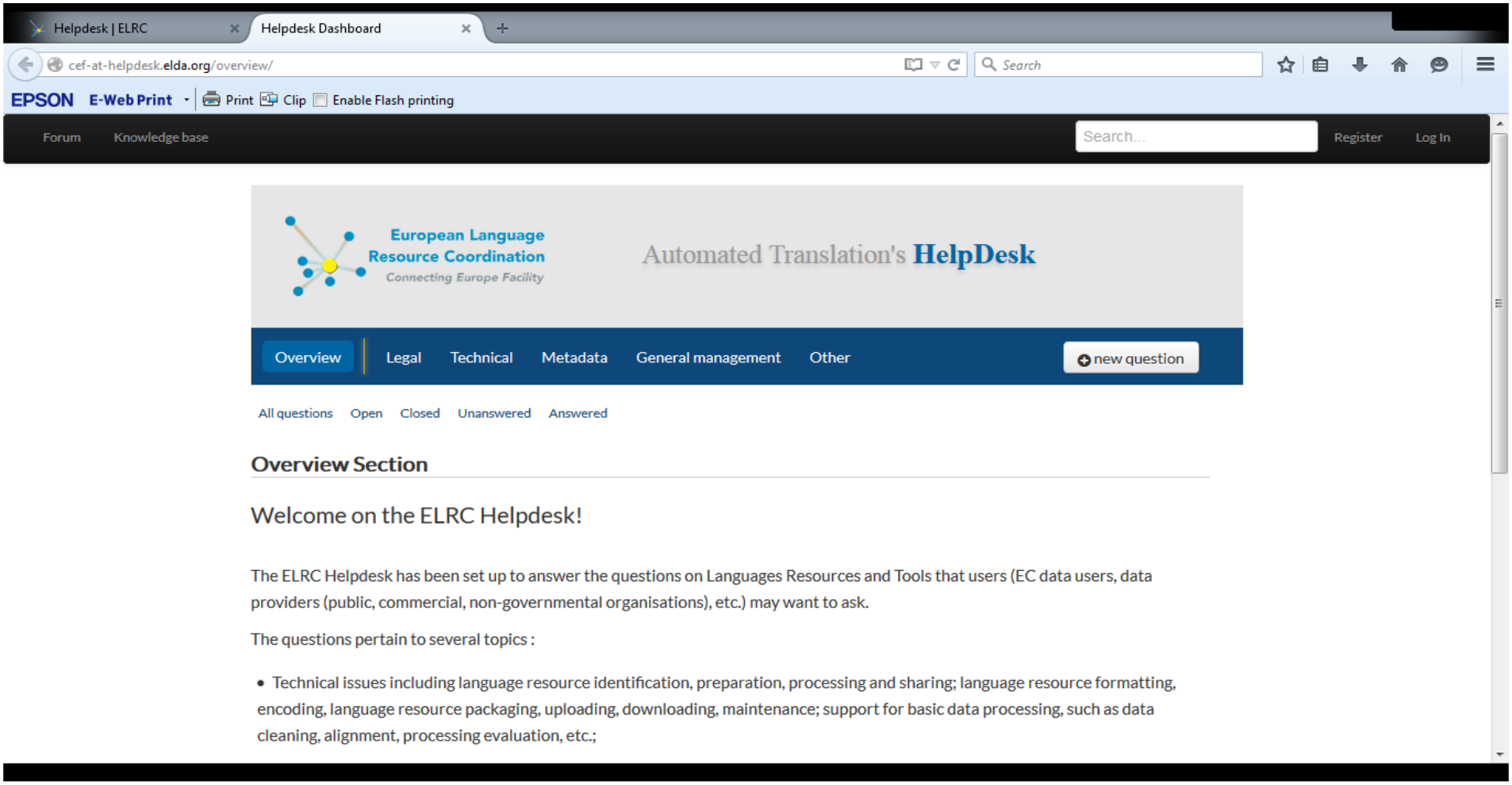
Ιστότοπος ELRC: Αποθετήριο



Ιστότοπος ELRC: Αποθετήριο



Ιστότοπος ELRC: Φόρουμ χρηστών



European Language Resource Coordination
Connecting Europe Facility

Συμπεράσματα

- 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