

# DATA COLLECTION APPROCHES



- A non-exhaustive list of typical datasets that will be sought by ELRC as provided in the tender specification, includes:
  - translation memories,
  - aligned parallel corpora,
  - comparable corpora,
  - monolingual corpora,
  - terminologies, lexica
  - grammars, etc.



- Anything that contains “words”, preferences for “sentences”, even for sentences expressed in multiple languages, e.g.
  - Reports,
  - Speeches,
  - Contents on web pages,
  - Brochures, etc.
- Bags of “words”, “sentences”, multiple bags



- The main public entities are those with a mission at the national, regional, local, cross-borders, cross-countries (bi-lateral, multi-lateral) but could be also international organizations with a European basis and mission. These cover:
- Presidents, Kings or Queens head of states offices,
- National or Federal Ministries (inc. inter-ministerial bodies) and their various departments,
- National Public bodies (Parliaments, Senates, Supreme Courts/Attorney General's Office, National Banks and their international branches if any or their representations in foreign countries),
- National representatives outside the country (Embassies, Consulates, Cultural and educational centers e.g. Cervantes, etc.),
- National and official language custodians (e.g. national language institutes, but also language centers in foreign countries e.g. Real Academia in Spain but also Cervantes or Goethe-Institut: outside their home country),
- **Regional “governmental” bodies and regional assemblies (e.g. Landers, States, Regions, Provinces, departments, districts, and other country's administrative regions and territorial entities as one can identify within each country),**
- Local authorities (Municipalities, county/city councils, and other city administrations, e.g. Twin-cities managing bodies, etc.),
- Cross-border political /economic /social bodies (Belgium-Netherlands joint association, Italian-Slovenian, North-South Catalan bodies, etc.) and other bilateral/multilateral public-like organizations,
- Political, social, sports, etc. organizations (political parties, unions and union federations, sports federations, etc.),
- International organizations with European basis (Transparency International, Human-Watch, UNICEF, OCDE, etc.),
- Consumer bodies and Conflict and Dispute resolution bodies,
- Security and emergency services, national and/or municipal police forces, and other law enforcement bodies, Fire services, etc. but also EU and International levels (Interpol),
- National Archives and Public Sector Information Management bodies (Etalab, Data Open Portals, National Libraries),
- Any organization that we will discover with production of “administrative” language material in particular bilingual/multilingual texts,

# What format is needed? Digital textual data





## Dublin Core Metadata Element Set

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

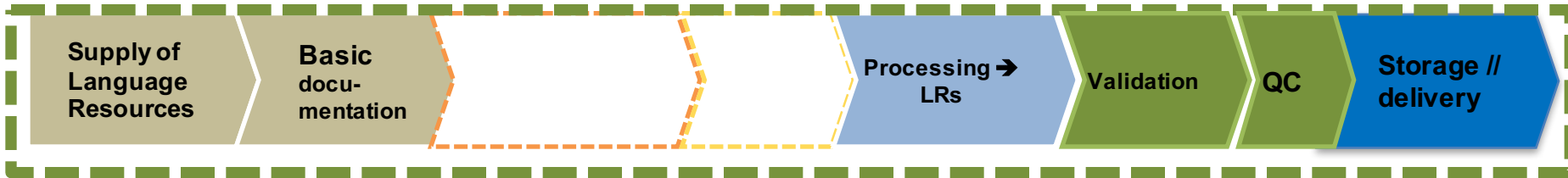
- On a case by case we can define what is ESSENTIAL:
  - Sources of data (trustability, quality, etc.)
  - Domain specific
  - Languages
  - Rights if not public



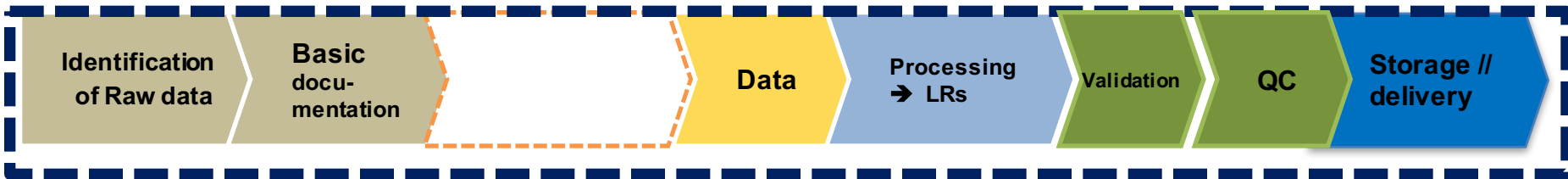
1. Sources coming from Identified partners (URLs)
  - URLs from Public Sector Bodies (“.html” style)
2. “Raw” Data coming from identified providers
  - Data formatted as .doc, .txt, .... .xml
3. Language Resources coming from “reliable providers”  
e.g. .tmx with identified languages, domains etc.



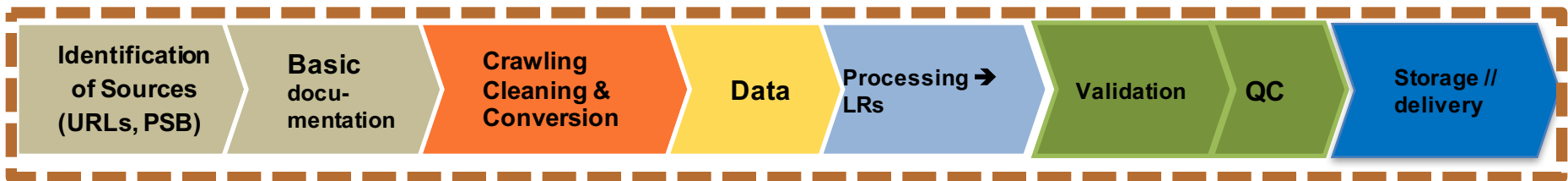
## 1. Starting from “LRs” ... RePurposing



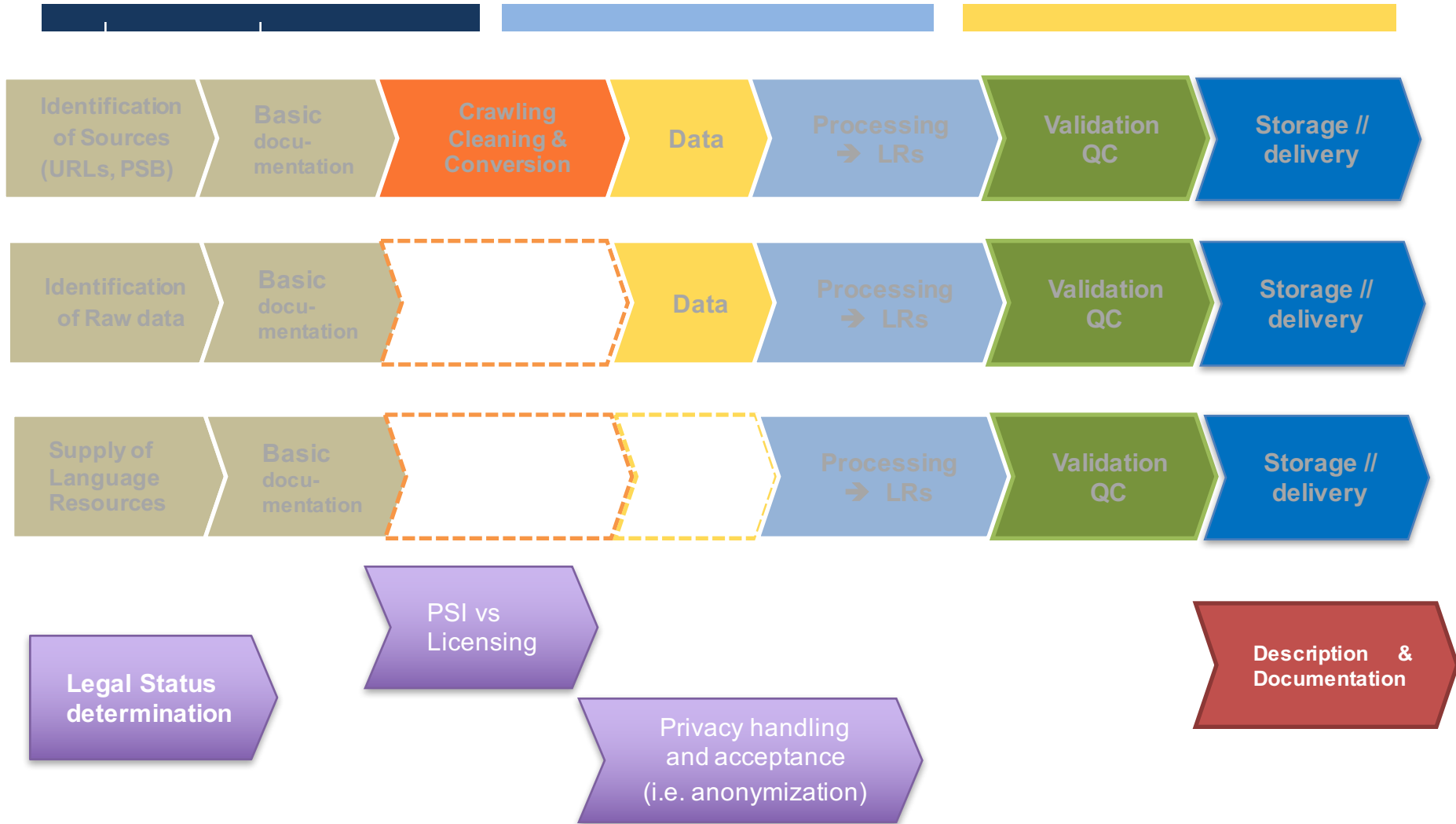
## 2. Starting from “Raw Data”



## 3. Starting from “Sources” (e.g. URLs)









## ➤ Scenario 1: From Language Resources

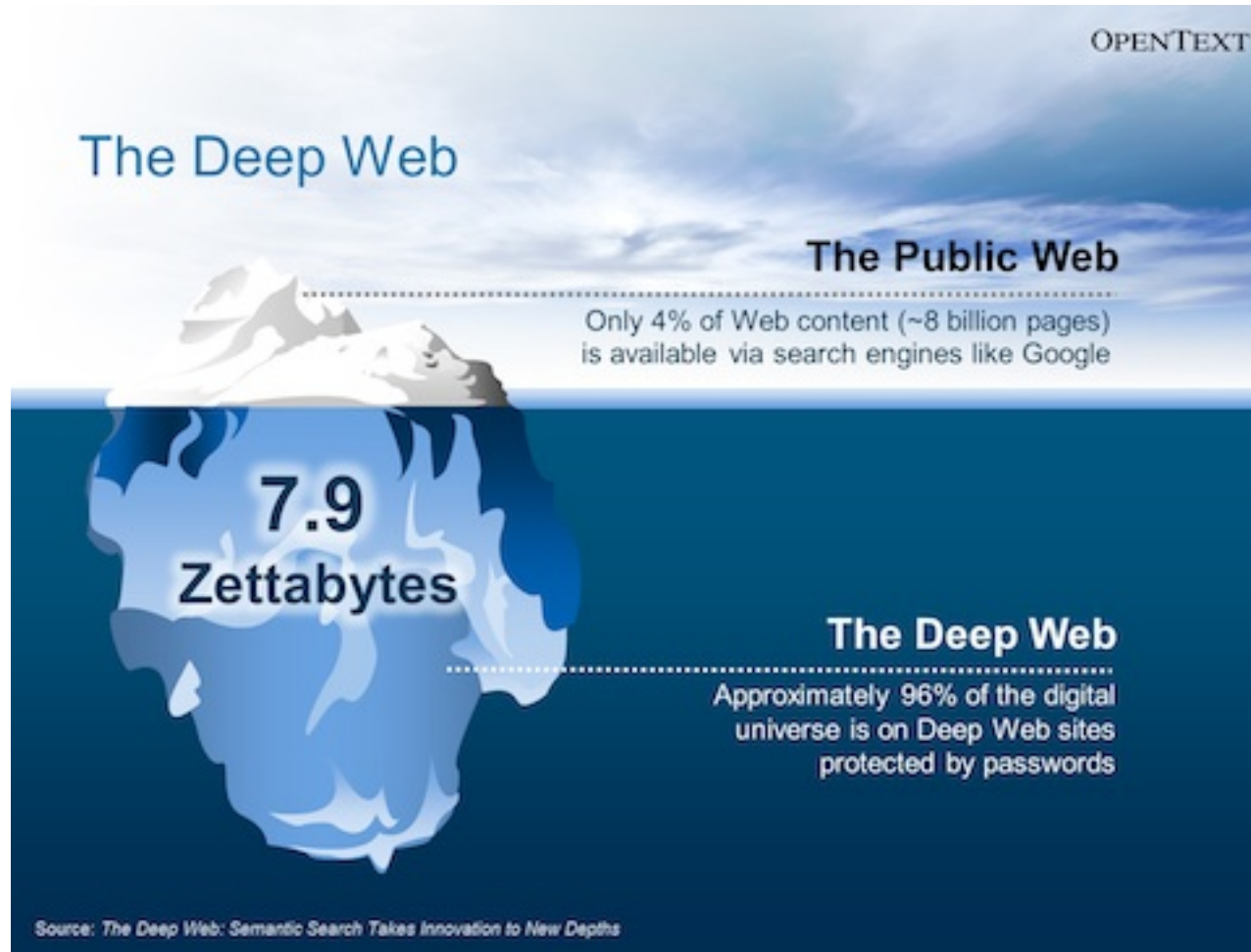
- See below; starting at conversion to LRs

## ➤ Scenario 2: From Data-set

- See below; starting at Data-set step

## ➤ Scenario 3: Select Sources and run the whole procedure:

- From Sources to Language Resources & Validate all steps (identification, qualification, crawling, cleaning/formatting, conversion to Data-sets, Processing, Validation, QC & assessment, conversion to LRs, depositing within CEF-share)







- Identification and qualification of sources
  - To browse the database <http://cef-at-sources.elda.org/>  
user name: inventory\_users  
password: FgOu8woYHUYr
  - See your login/password sent by email



## 1. Task 1: Identification

### i. Identification of sources (URLs from PBS)

- See the ELDA CEF – Database

### ii. Identification of Raw Data

- .doc, .pdf, ... identification of providers (URL if any)

### iii. Identification of Language Resources

- e.g. .tmx (providers, languages, domains, etc.)



- Task 2: Preliminary Documentation
  - i. Identification of sources (URLs from PBS)
    - See the ELDA CEF – Database
  - ii. Identification of “donated” Raw Data
    - .doc, .pdf, ... identification of providers (URL if any)
  - iii. Identification of donated Language Resources
    - e.g. .tmx (providers, languages, domains, etc.)



## • Task 3: Collections

### i. Selection of sources to harvest

- Define crawling processes, e.g. terms/Domains, “stop” arguments, etc.
- Crawling /Storage/Cleaning/Alignment if ML
- See the ELDA Storage site (<https://cef-at.elda.org/LR1>; LR2; etc.

### ii. Selection of “donated” Raw Data for processings

- Converting .doc, .pdf, ... to usable formats (? .xml)
- Storage/Cleaning/Alignment if ML
- See the ELDA Storage site (<https://cef-at.elda.org/LR1>; LR2; etc.

### iii. Storage of donated Language Resources

-





- Task 4: Metadata & documentation
  - i. Identification of the meta-data elements for Crawled data
    - Filling CEF-SHARE (@ILSP)
  - ii. Identification of the meta-data elements for the processed “donated” Raw Data
    - Filling CEF-SHARE (@ILSP)
  - iii. Identification of the meta-data elements for the donated Language Resources
    - Filling CEF-SHARE (@ILSP)

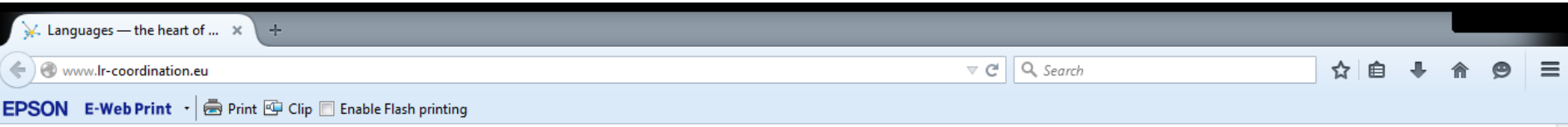


- Task 5: QC & Validation Criteria\*
  - Validation Criteria Definition
    - As by the EC // Consortium Document
  - Validation of meta-data elements
    - As by the Inception report
  - Validation of Language Resources
    - i. for Crawled data
    - ii. for the processed “donated” Raw Data
    - iii. for the donated Language Resources



- Task 6: Various Storage mechanisms
  - Storage of data-sources (ELDA-dbb)
  - Storage of Language Resources Metadata elements
  - Storage of Raw Data (.doc, .xml, etc.)
  - Storage of final validated LRs
  - What is stored where:
    - CEF-SHARE
    - CEF-Repository
    - DGT-Repository

# Helpdesk and Support

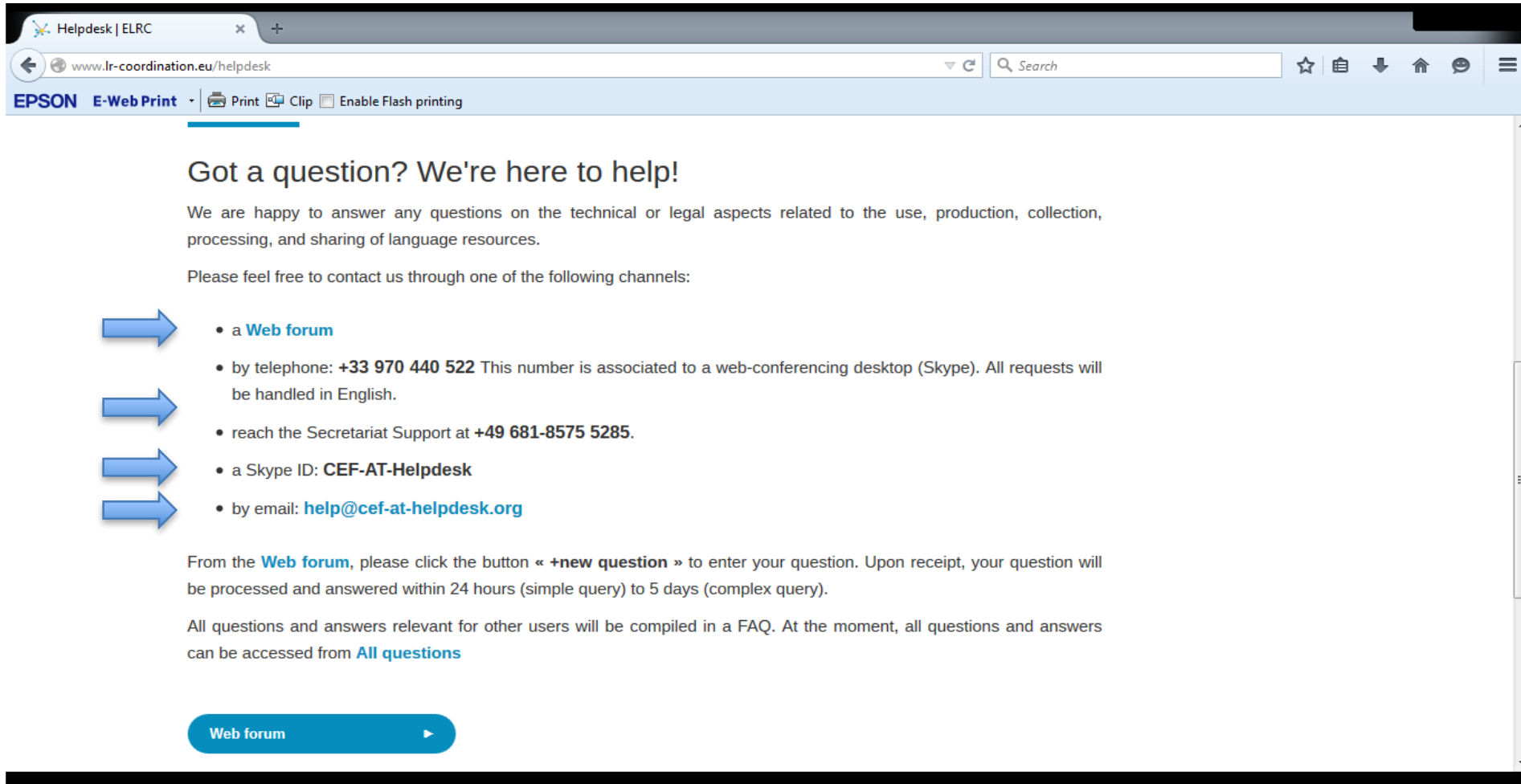


[Home](#) [About](#) [News](#) [Helpdesk](#) [Events](#) [Resources](#) [Anchor Points](#) [Multilingual Europe](#)

European Language  
Resource Coordination



Languages — the heart of  
Multilingual Europe



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.

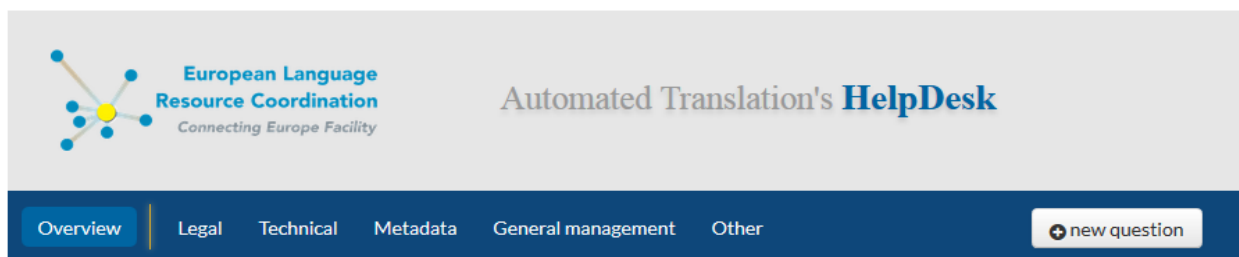
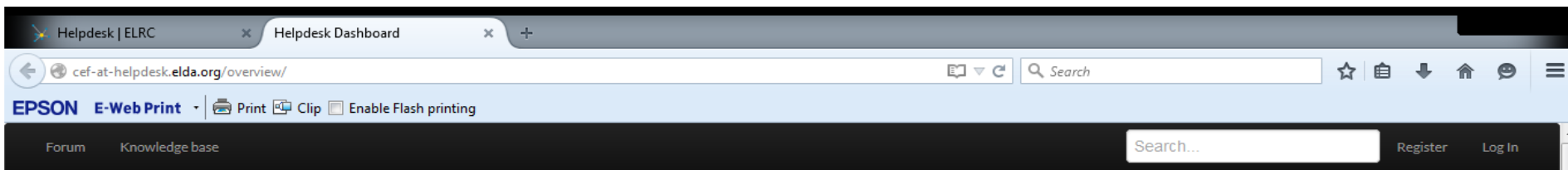
Please feel free to contact us through one of the following channels:

- 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](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**

[Web forum](#)



[All questions](#) [Open](#) [Closed](#) [Unanswered](#) [Answered](#)

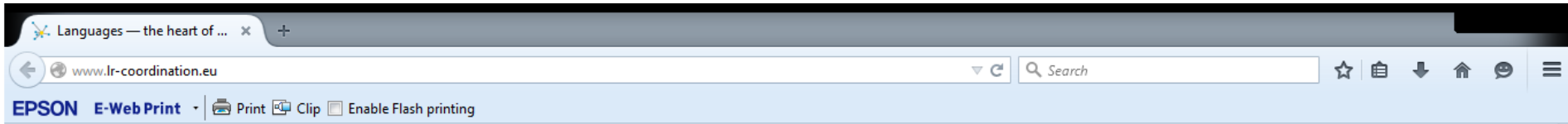
## Overview Section

### 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.;



[Home](#)

[About](#)

[News](#)

[Helpdesk](#)

[Event](#)

[Resources](#)

[Anchor Points](#)

[Multilingual Europe](#)

European Language  
Resource Coordination



Languages — the heart of  
Multilingual Europe





**European Language  
Resource Coordination**  
*Connecting Europe Faculty*

## Automated Translation's Addition of Data Sources

Data Sources are identified websites URLs that could be exploited, through a crawling process, for the preparation of Language Resources within the CEF.AT platform, in particular parallel corpora to be built up from multilingual websites.

Please fill in the form below with any exploitable sources or other information on potential Language Resources.

**URL of the source\***

**Name of the source\***

**Comments on the source**

**Contact name**

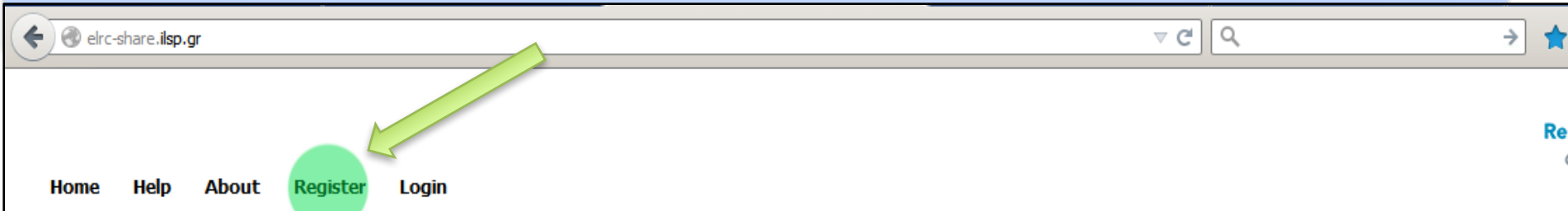
**Contact institution**

**Contact e-mail**

Inventory of relevant sources



- ELRC Repository



- Click "*Register*"

