European Language Resource Coordination
Connecting Europe Facility

Capitalize on your data

Best Practices for the future
Open Issues on how to contribute data

Khalid Choukri (ELDA)

European Language Resource Coordination
Connecting Europe Facility

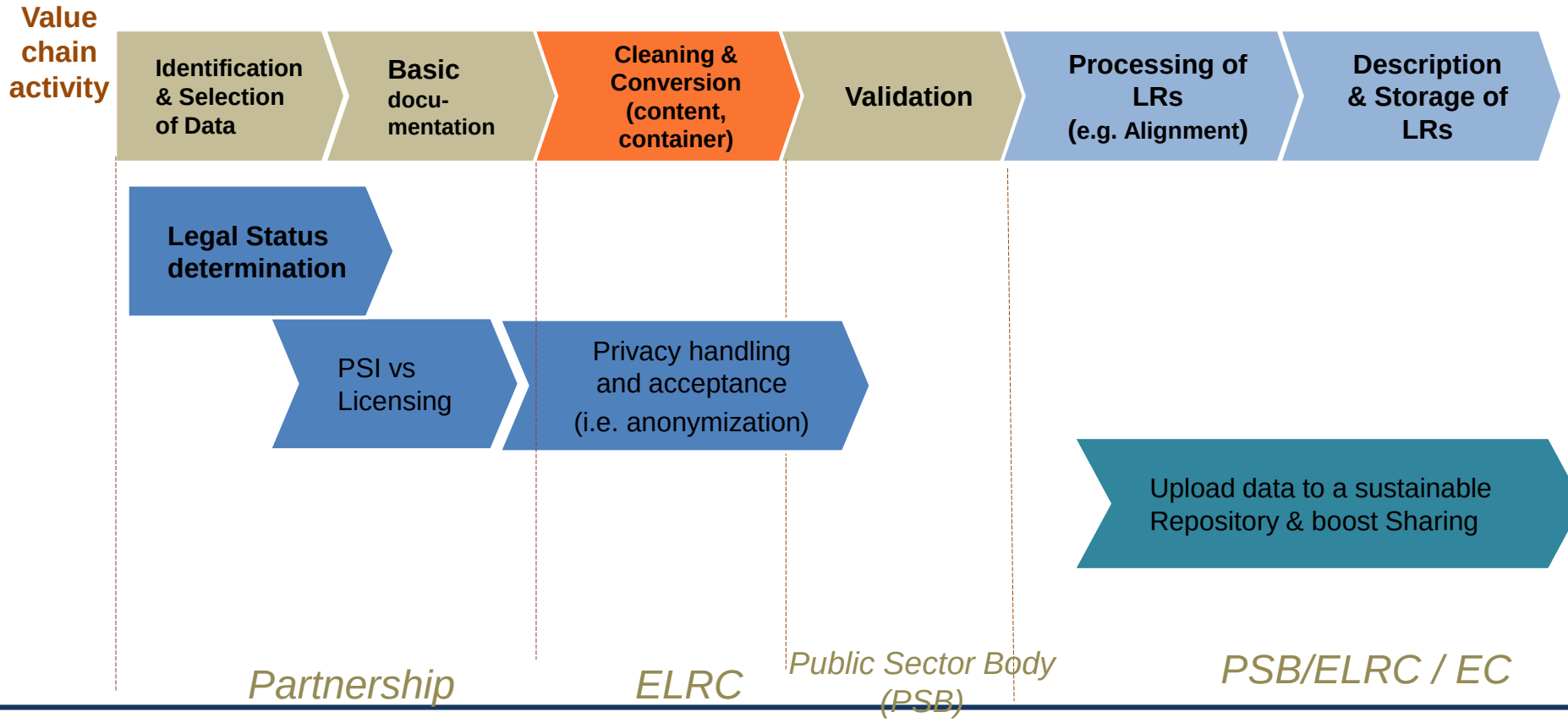
Overview

- We have seen the importance of data for Automated Translations
  - 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
- What can be done for the future to capitalize on the data assets
  - Our experience with Data Management Plans
  - Sustainable sharing

European Language Resource Coordination
Connecting Europe Facility

Illustration of data packaging workflow

Existing Data ➔LRs (Language Resources)



European Language Resource Coordination
Connecting Europe Facility

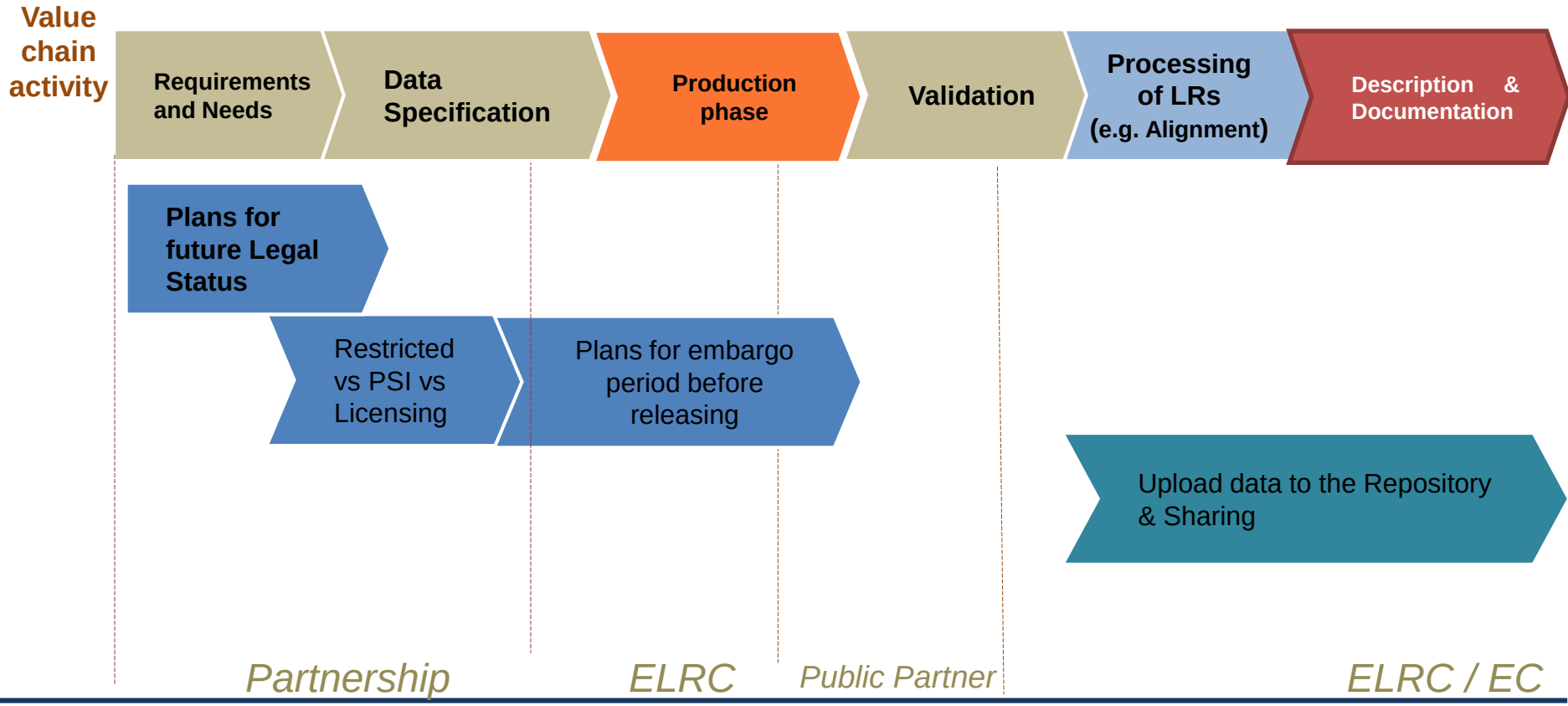
If new data
some best practices for data management

1) Analyse all phases of data development
2) Based on 1), create a data management plan
   - Legal, data workflow, formats, publication as PSI, …
   - **Relations with subcontractors and other partners**
3) Consider data sustainability
   - Data specification, production, validation, sharing & distribution, maintenance & preservation
4) Use the Web as an additional publication channel (see how ELRC can help)

European Language Resource Coordination
Connecting Europe Facility

Illustration of data packaging workflow

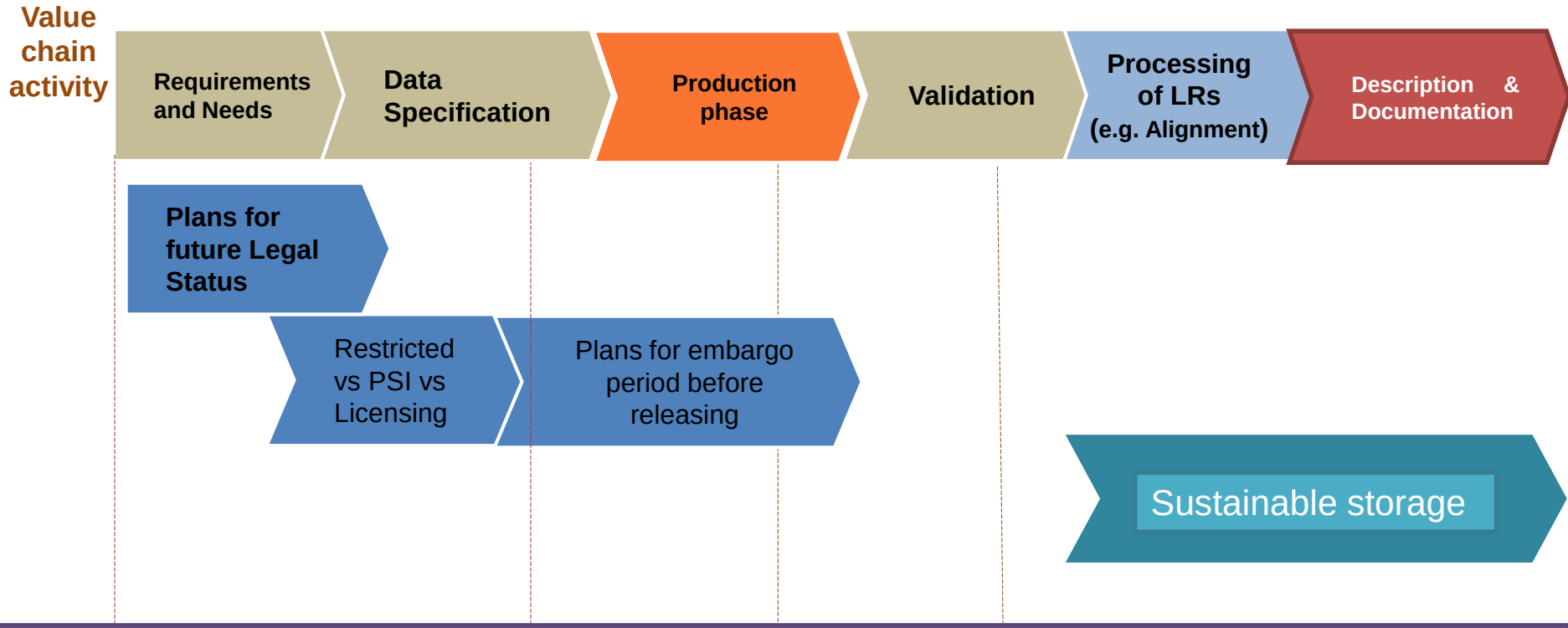
New Data ➔LRs (Language Resources)



European Language Resource Coordination
Connecting Europe Facility

Illustration of data packaging workflow

New Data ➔LRs (Language Resources)



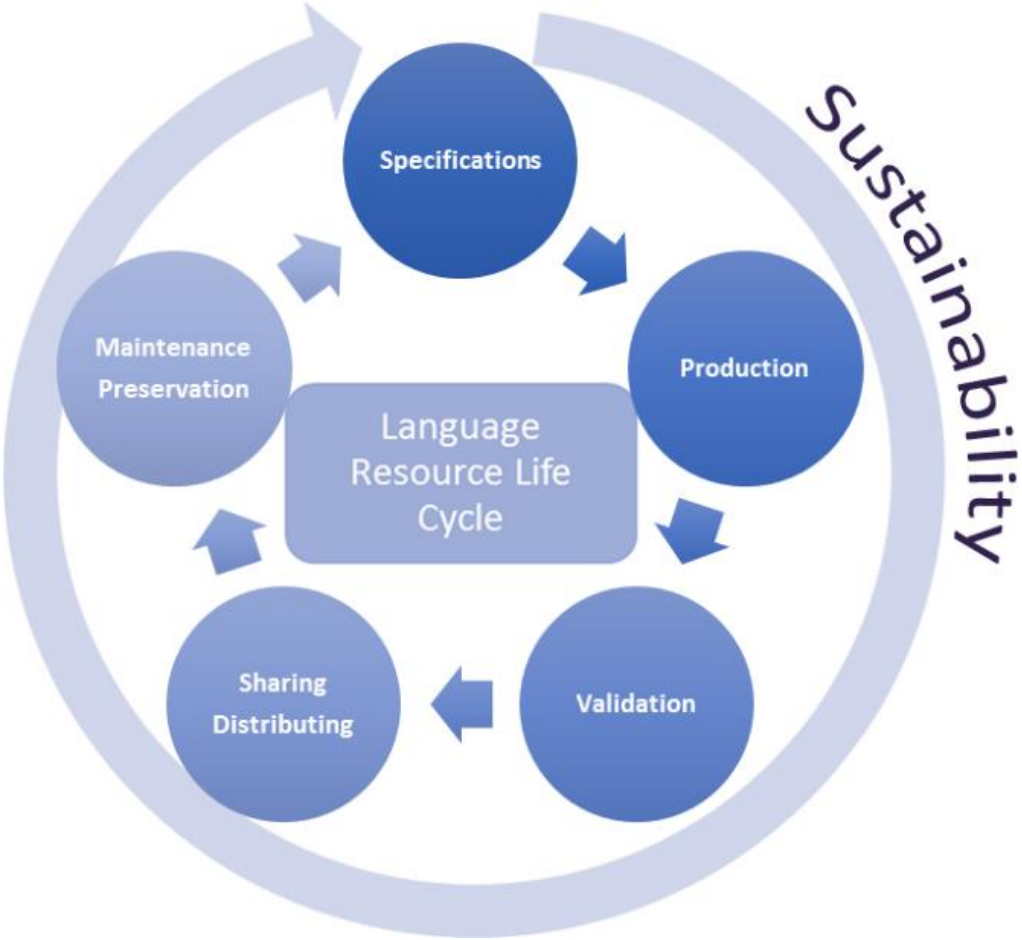
This can be part of the data management plan (DMP)

Concerns in creating a DMP

- Anticipate all potential legal issues
  - Ensure that your data IPRs are cleared
  - Ensure that the producing parties adhere to your right “ownership” (e.g. relations with LSP: ensure you keep all rights)
  - Ensure that all produced intermediary documents are yours (e.g. Translation Memories)
  - Check the privacy issues in advance and plan for anonymization if necessary
- Define your management plan with respect to the task
  - This has to account for the main goal (e.g. document writing, doc translation, etc.)
- Plan for repurposing (from documentation to LRs)
  - Request data in a usable format (not only PDFs but also TMX/Word/XML/TXT)
  - Make sure that your data uses up-to-date medium (no CDs?)
- Foresee for future publication and sharing as Public Sector Information (PSI)

European Language Resource Coordination
Connecting Europe Facility

Key elements of a Data Management Plan



Key elements of a Data Management Plan

- Specifications
  - Ensure that the original documents are described
  - Ensure that your needs are described
  - Anticipate what you can get as valuable resources (a side effect)
- Production
  - Whether internal or outsourced, check that the tools used are compatible with your needs and beyond (e.g. CAT, MT, etc.)
  - Ask for the list of tools and production software
  - Check if you can get texts in the multiple languages aligned to each other
  - Keep a clear documentation of the data being produced (meta-data)

Key elements of a Data Management Plan

- Validation
  - In addition to your quality control, you may want to use some of the validation tools (lexical coherence, syntactic analysis, etc.)
- Sharing/distribution
  - Ensure your data falls within the PSI directive as transposed in your country
  - If not, foresee an open and permissive licence
  - If privacy is an issue, plan necessary procedures to handle these
- Maintenance/preservation
  - The best option is often to partnership with a data centre
  - See how ELRC can assist you
  - There is also the “option” of national open data portal
  - Only “putting” data on the web is not a sufficient option (referencing?)

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 indicated 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. often is not possible to outsiders
- 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

Conclusions

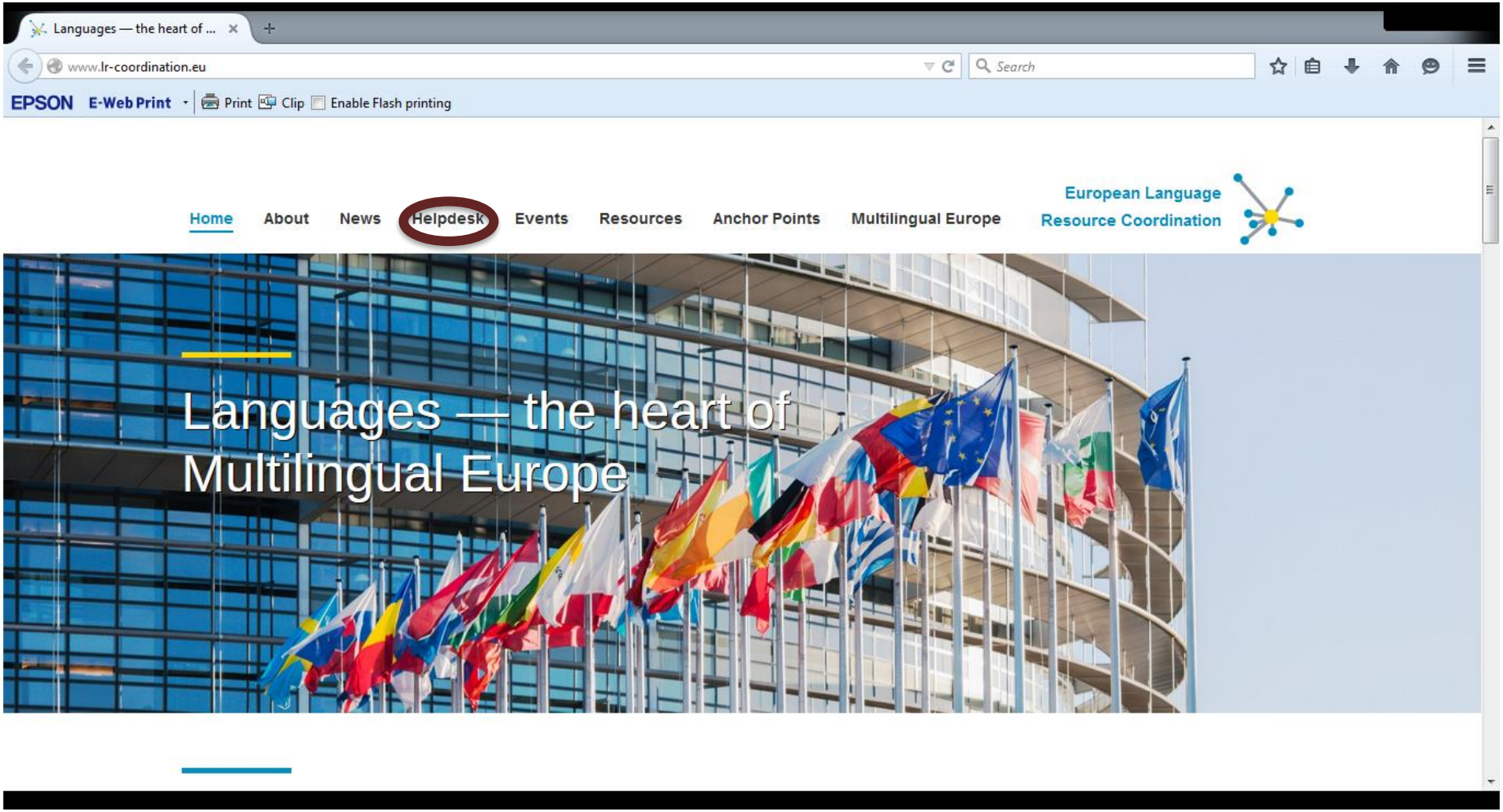
- Repurposing existing data (human translations) is the best way to improve Automated Translation quality
- Data-driven paradigms provide an efficient way to leverage value from existing resources
- ELRC can help reviewing data for suitability (at any phase)
- Do not underestimate the value of your language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

http://cef-at-sources.elda.org/add_source/

Helpdesk and Support

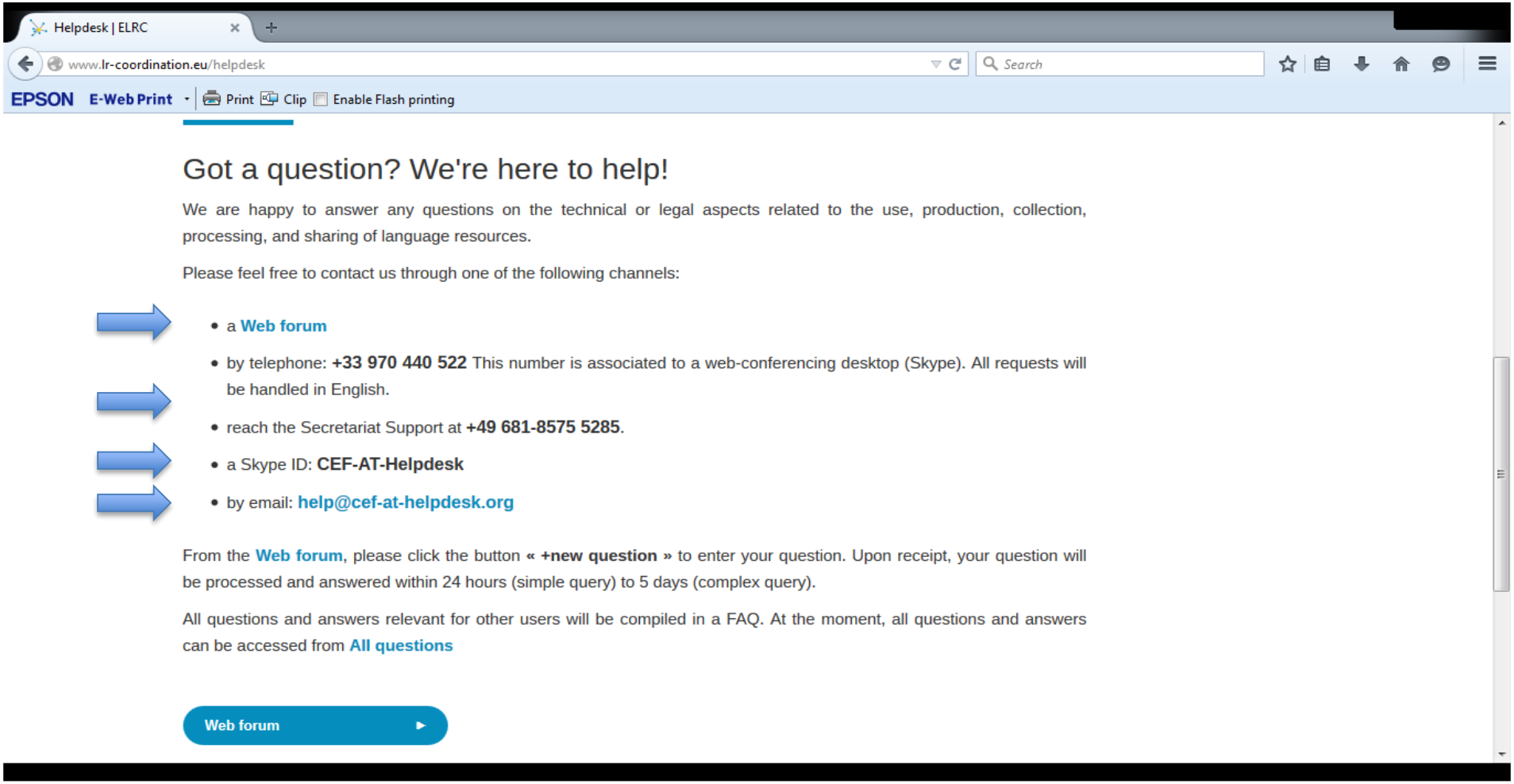
European Language Resource Coordination
Connecting Europe Facility

ELRC Portal
www.lr-coordination.eu



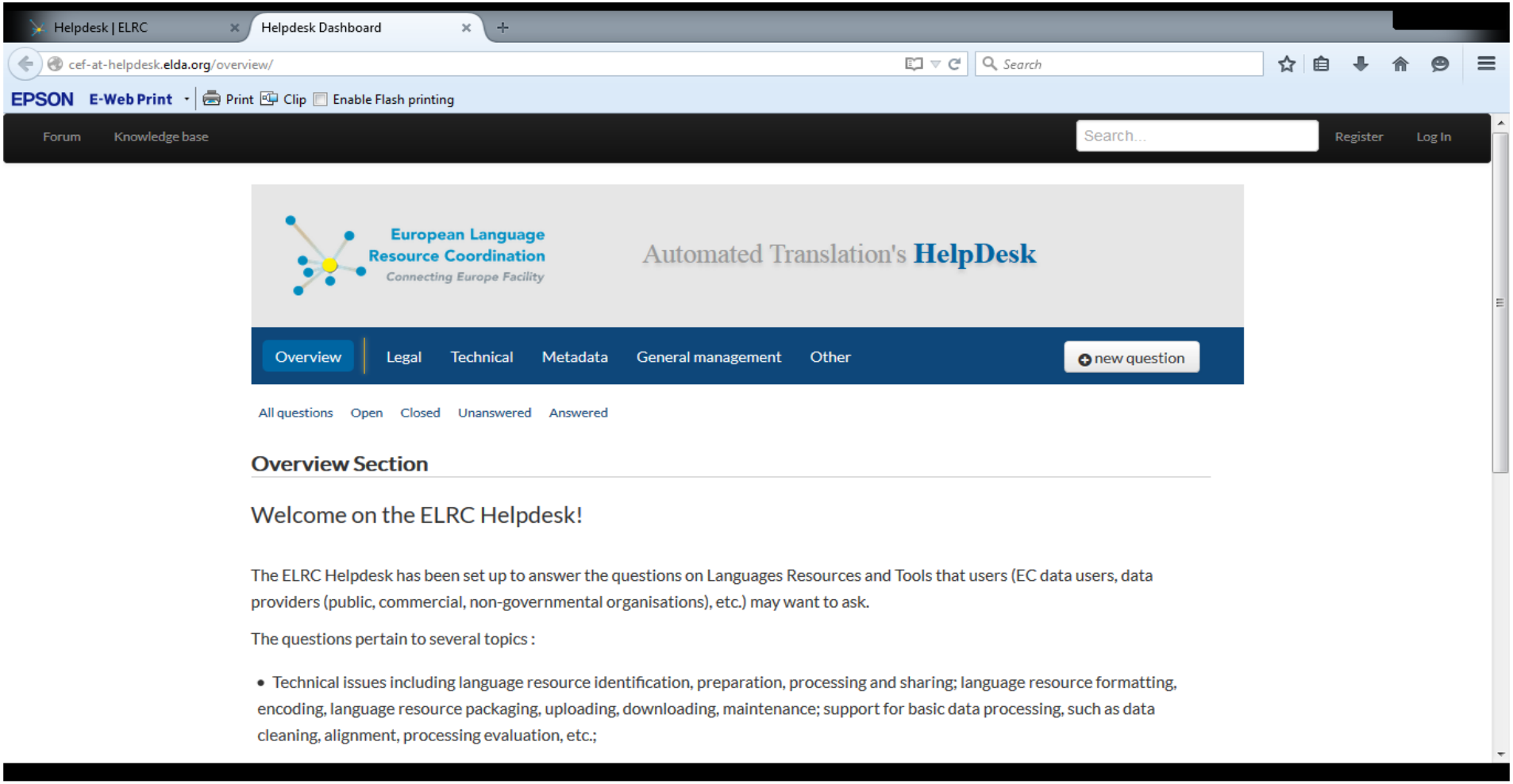
European Language Resource Coordination
Connecting Europe Facility

ELRC Portal: Helpdesk



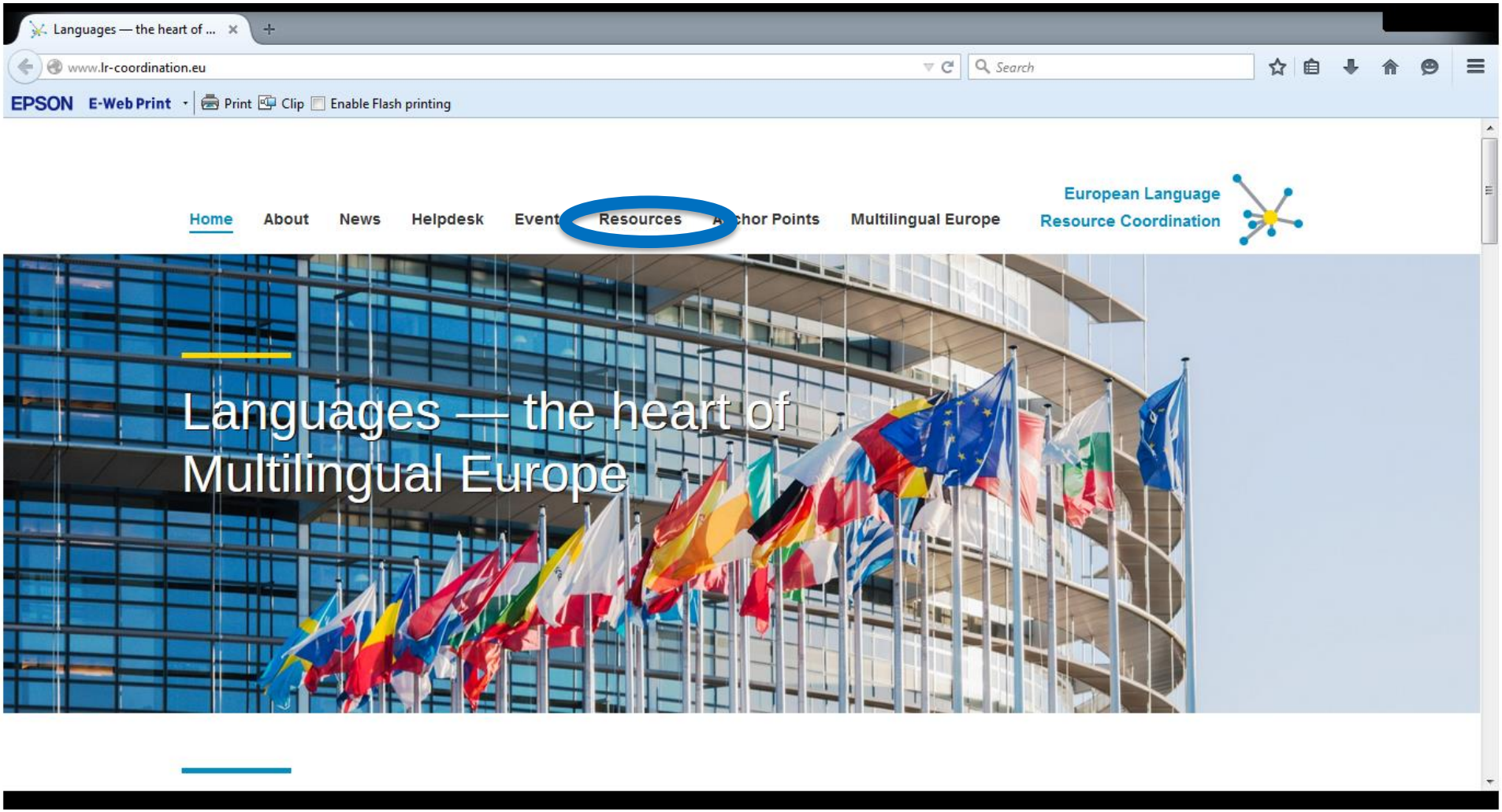
European Language Resource Coordination
Connecting Europe Facility

ELRC Portal: Web Forum

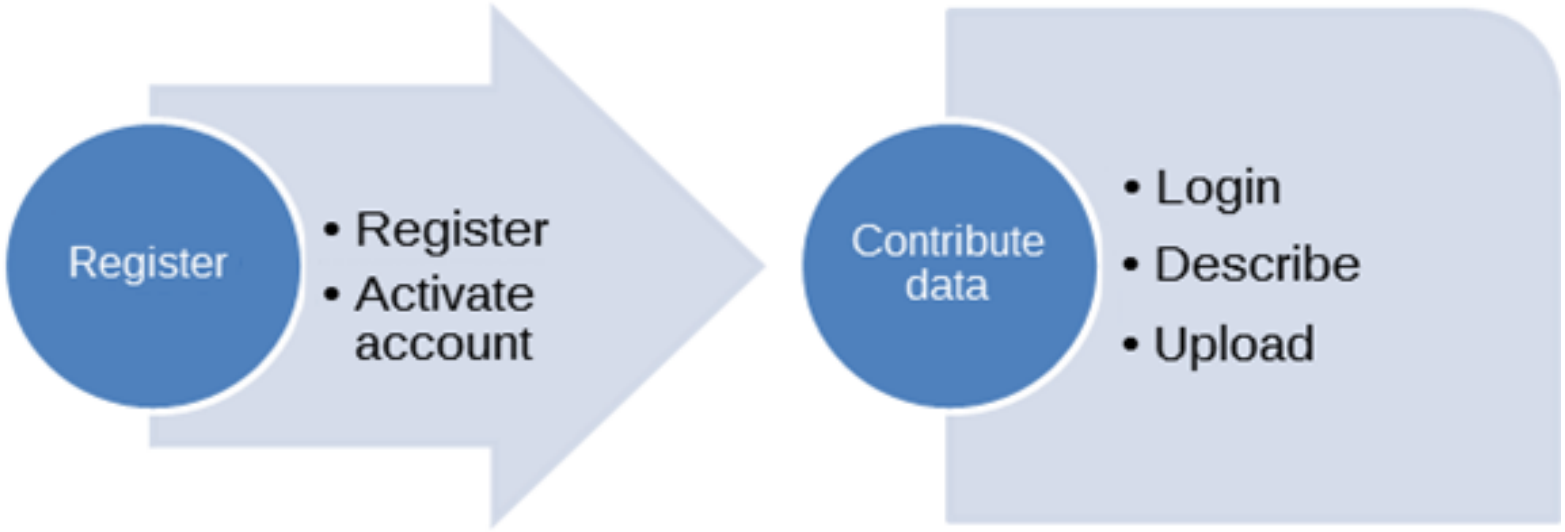


European Language Resource Coordination
Connecting Europe Facility

ELRC Portal
www.lr-coordination.eu

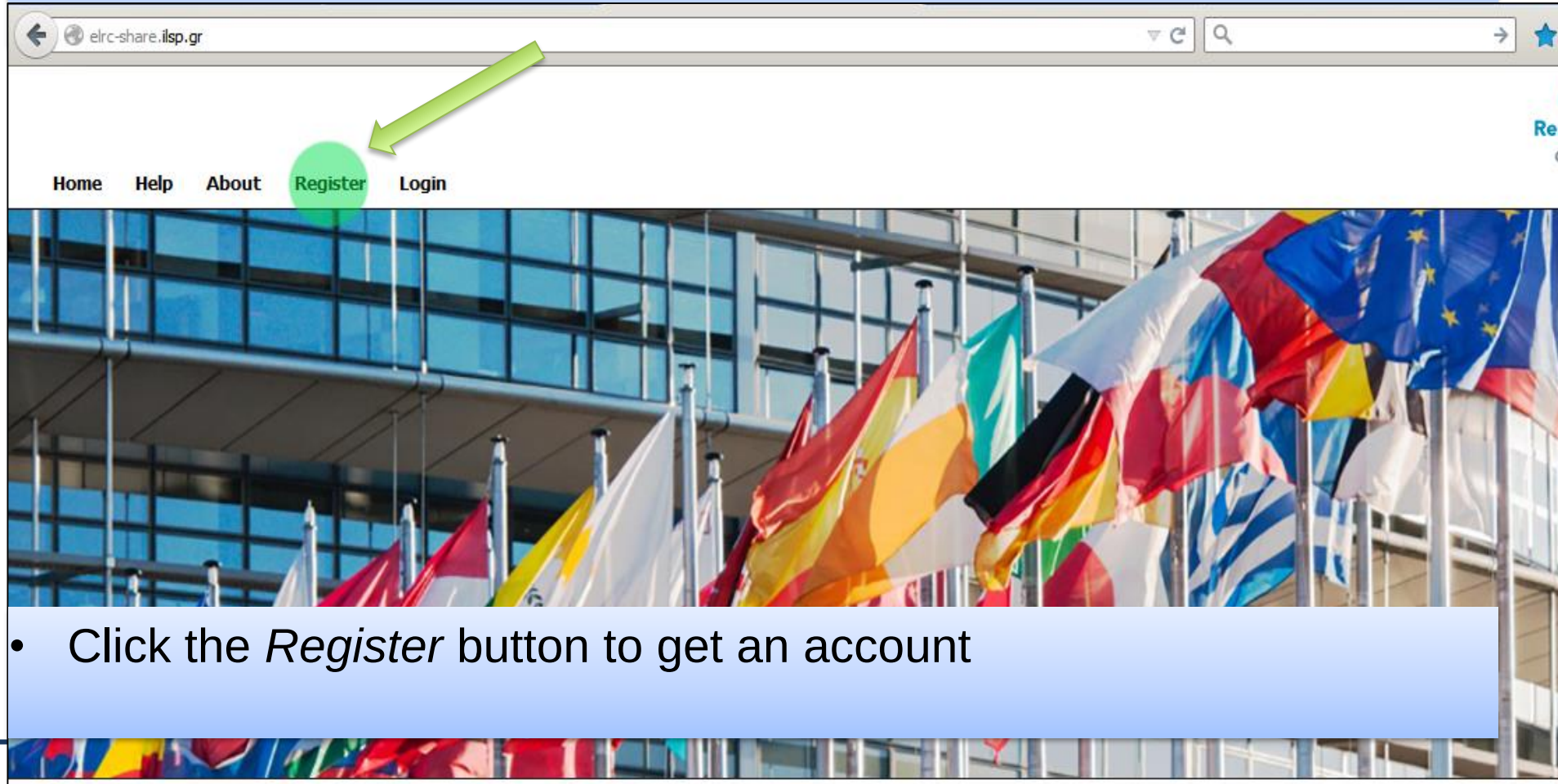


The resources part



How to Contribute Language Resources (1/7)

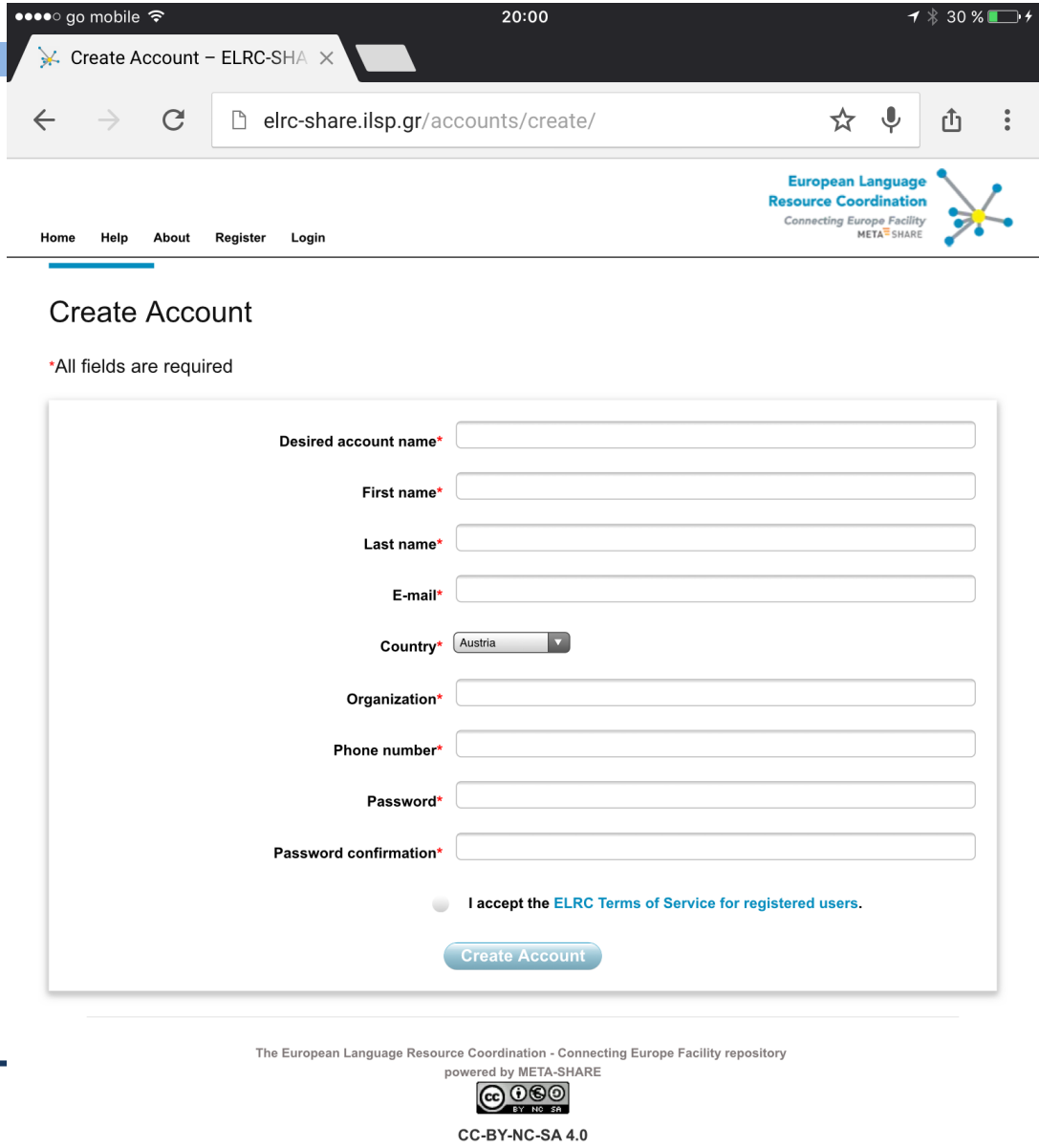
- Go to the ELRC Repository (through the RESOURCE Link)



- Click the *Register* button to get an account

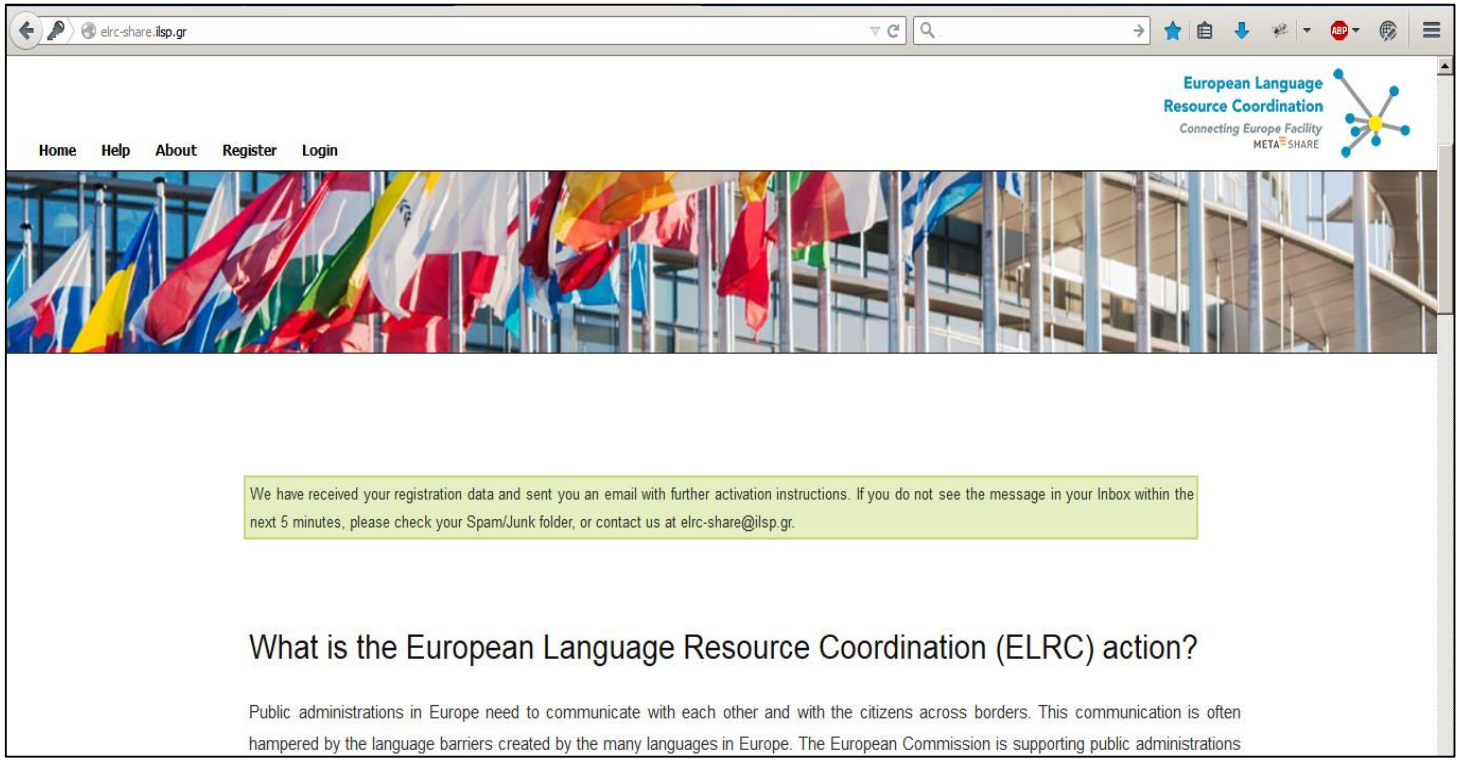
How to Contribute Data (2/7)

- Fill in the info
- Read the *Terms of Service* and click *Accept* if you agree
- Click the *Create Account* button



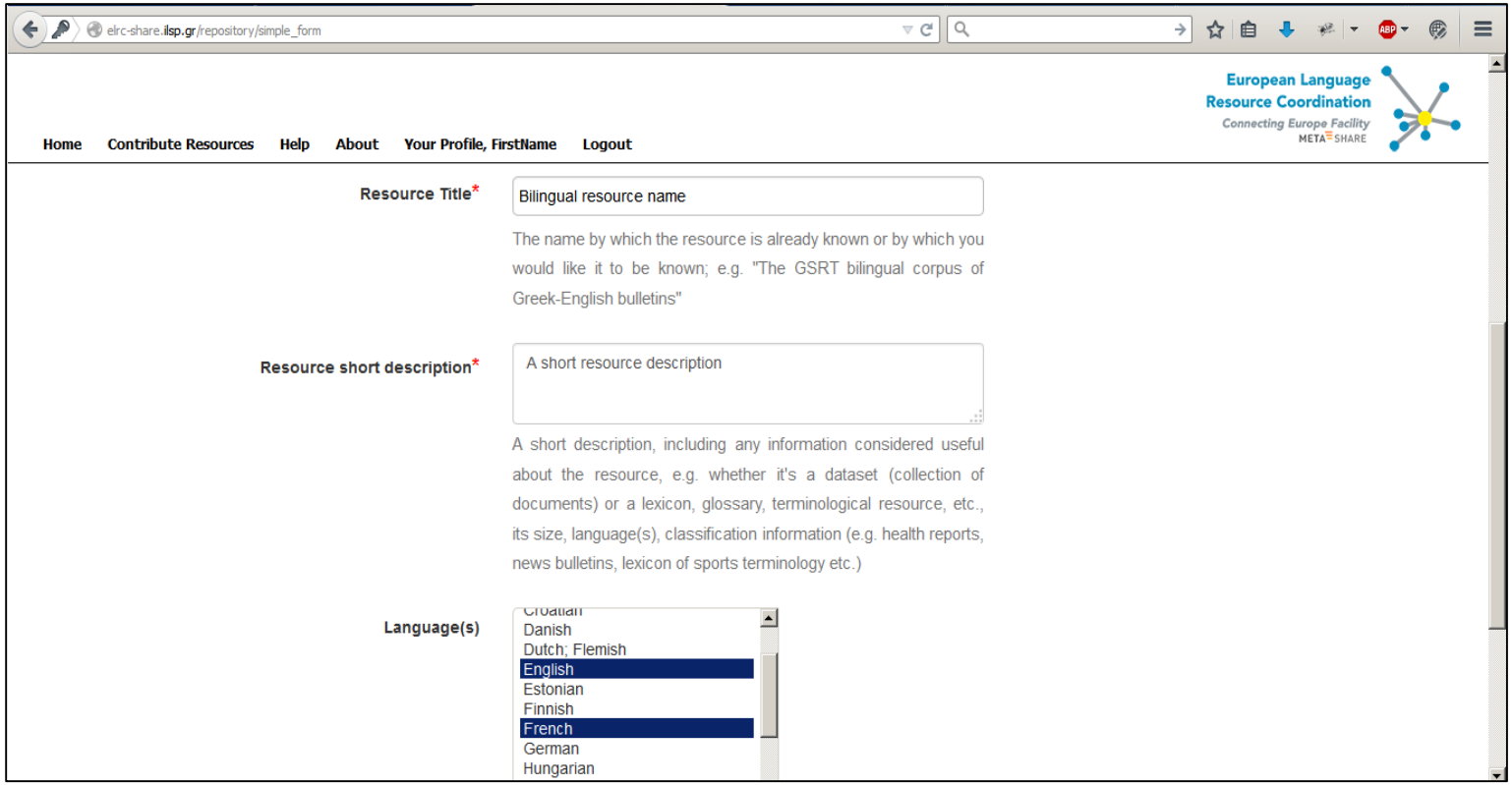
How to Contribute Data (3/7)

- Your request is acknowledged and an activation email is sent to the address you indicated
- Check your email and click the activation link



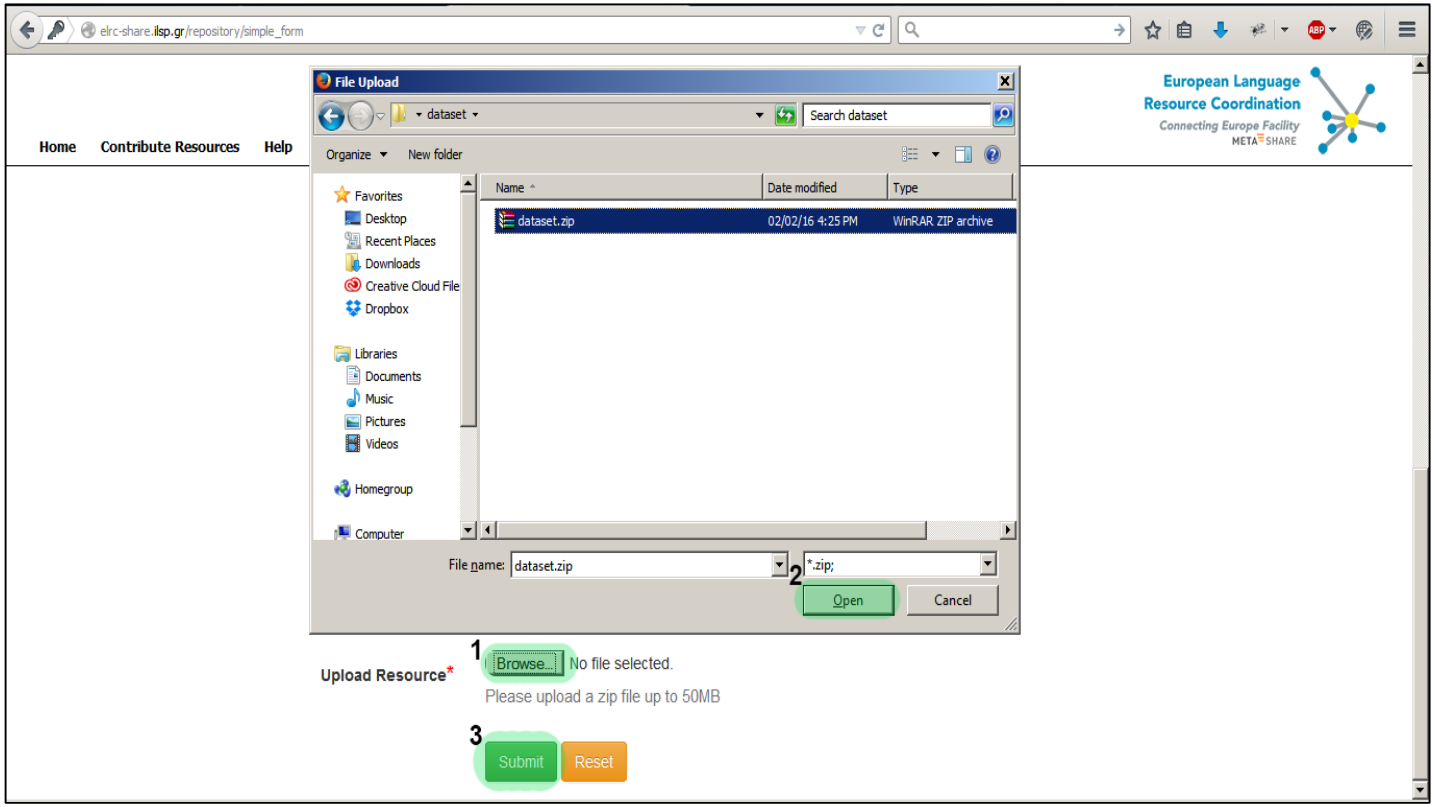
How to Contribute Data (5/7)

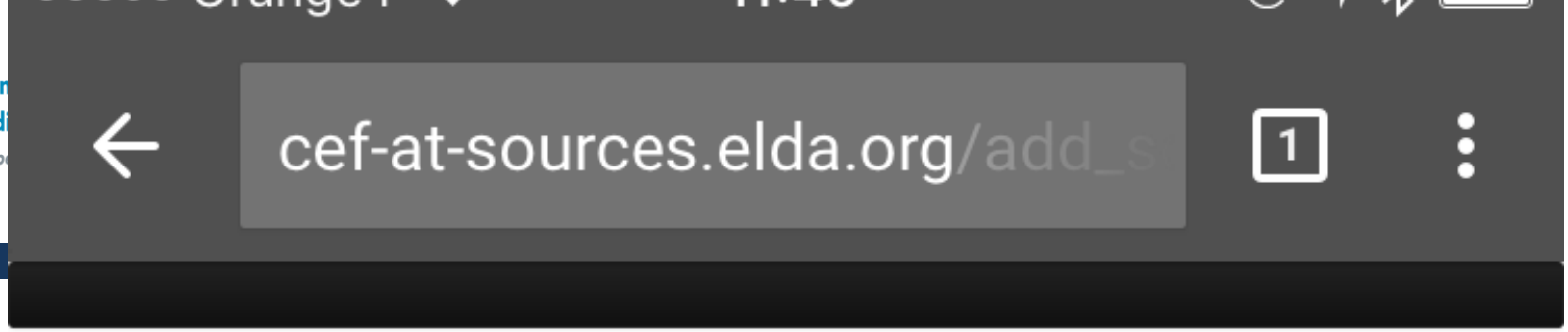
- Fill in the details of the dataset



How to Contribute Data (6/7)

- Browse your computer for the respective .zip file containing your data
- Click *Submit*





European Language Resource Coordination
Connecting Europe Facility

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*

URL of the source

Name of the source*

Name of the source

Comments on the source

Comments on the source

Contact name

Contact name

Contact institution

Contact institution

How to Suggest SOURCES of data