

Capitalize on your data

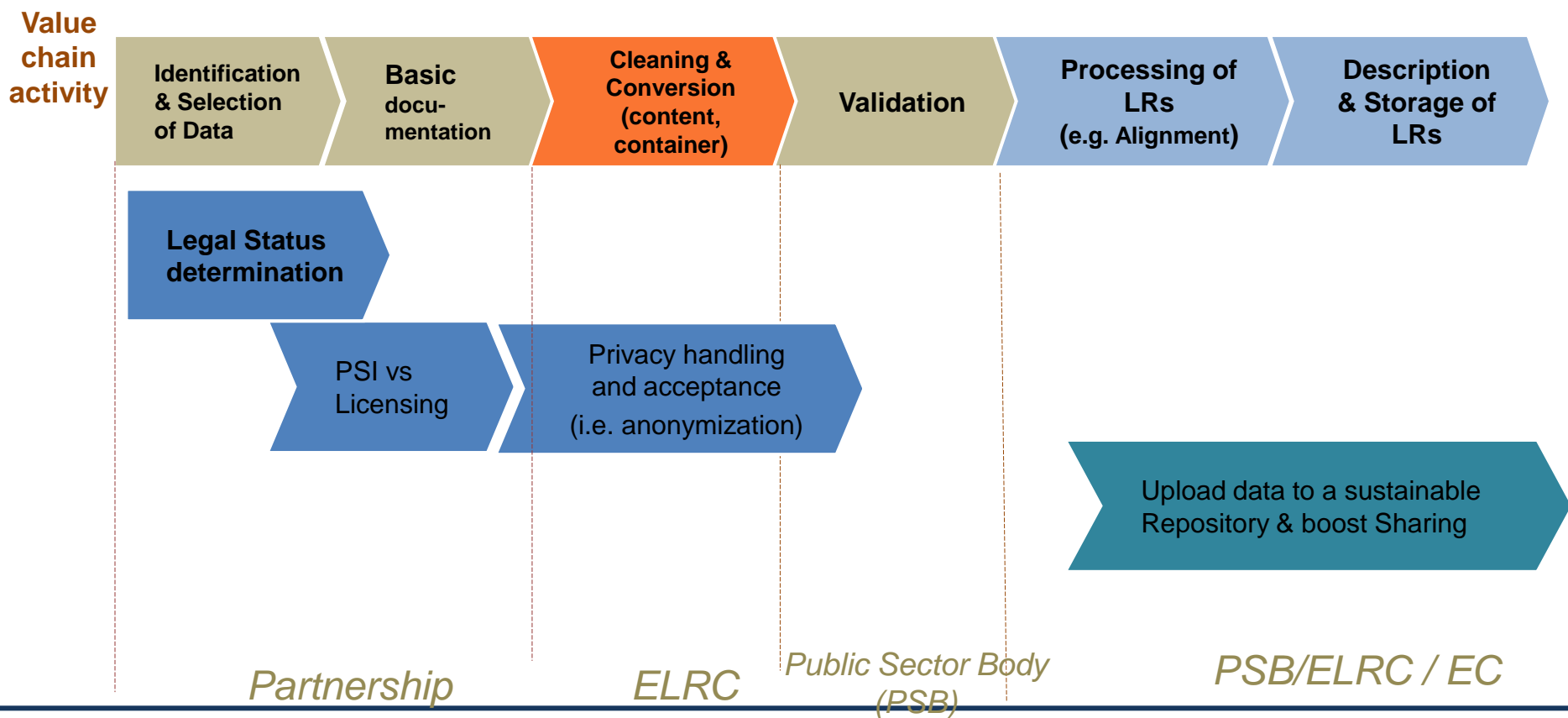
**Best Practices for the future
Open Issues on how to contribute data**

Khalid Choukri (ELDA)



- 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

Existing Data → LRs (Language Resources)

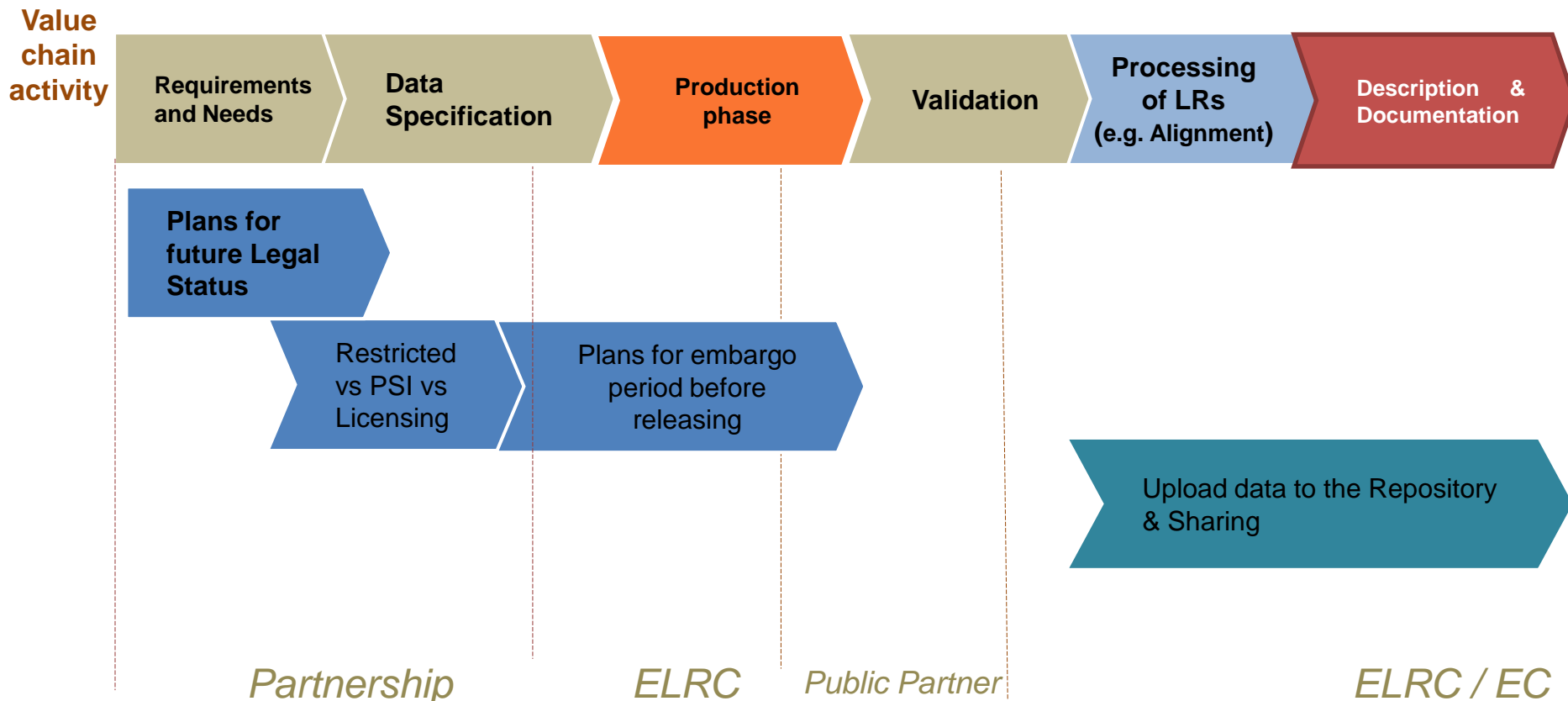




- 1) Analyse all phases of data development
- 2) Based on 1), create a data management plan
 - Legal, data workflow, formats, publication as PSI, ...
 - **Relations with subcontractors and other partners**
- 3) Consider data sustainability
 - Data specification, production, validation, sharing & distribution,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

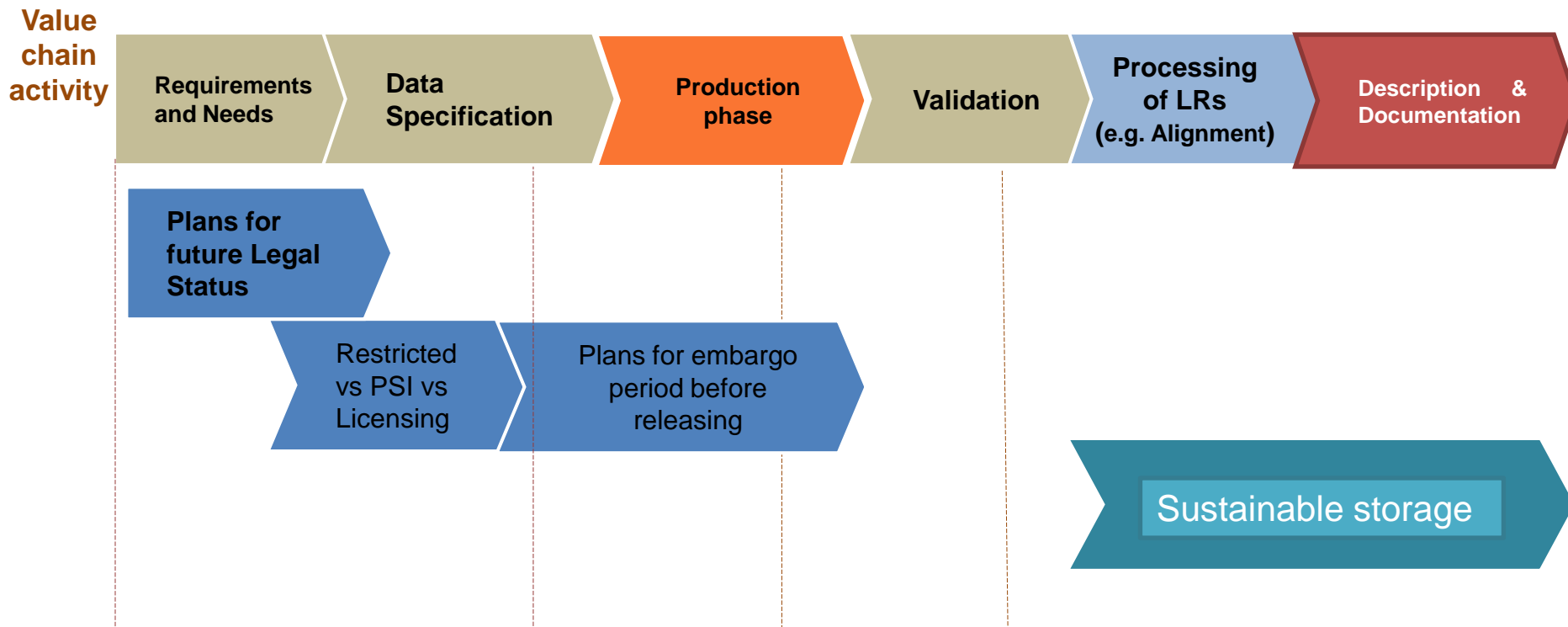


New Data → LRs (Language Resources)





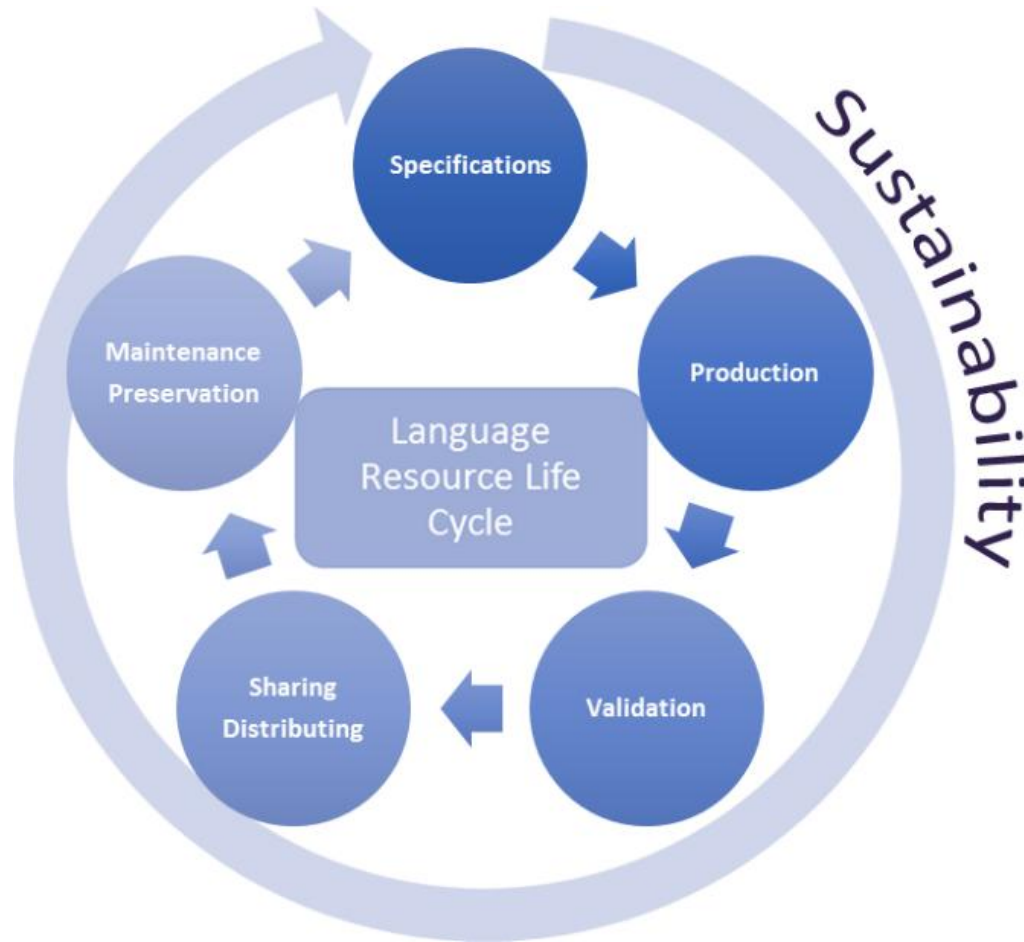
New Data → LRs (Language Resources)



This can be part of the data management plan (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)





– 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)



– 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?)

- 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

- 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.



- 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](#)

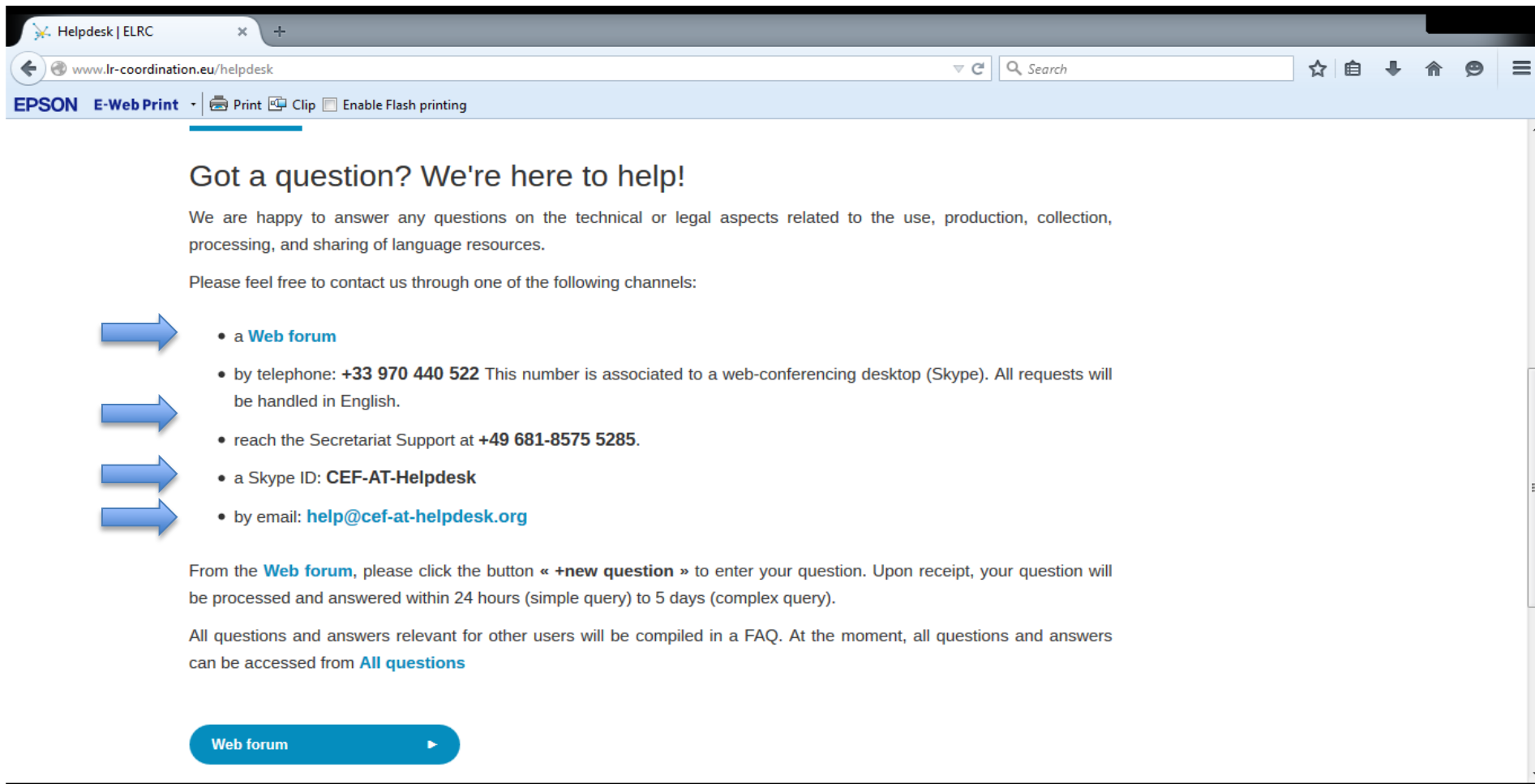
[Anchor Points](#)

[Multilingual Europe](#)

European Language
Resource Coordination



Languages — the heart of
Multilingual Europe



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.

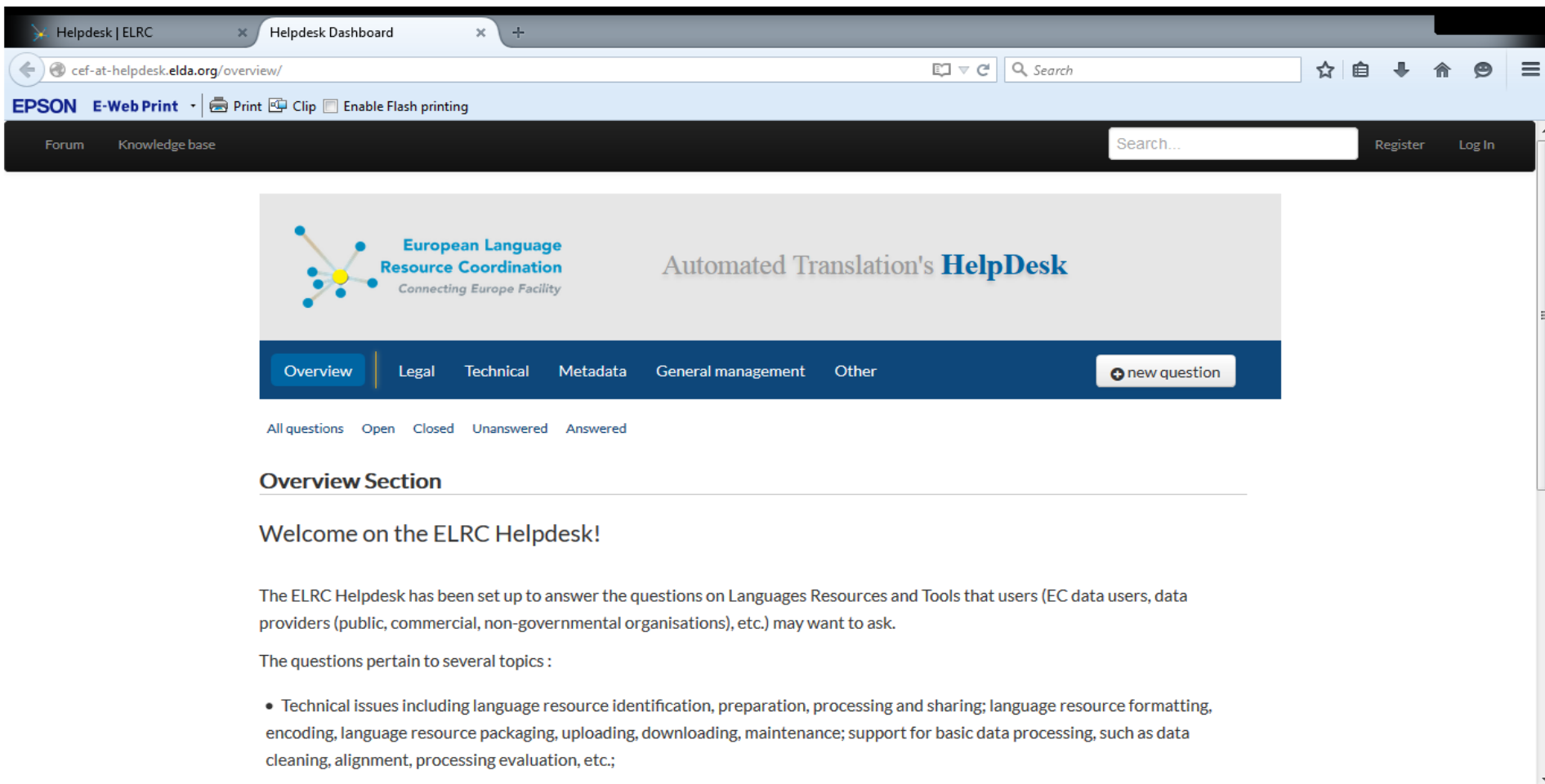
Please feel free to contact us through one of the following channels:

- 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 features a dark navigation bar with 'Forum' and 'Knowledge base' links, a search bar, and 'Register' and 'Log In' buttons. Below this is a header section with the ELRC logo and the text 'Automated Translation's HelpDesk'. A dark blue navigation bar contains tabs for 'Overview', 'Legal', 'Technical', 'Metadata', 'General management', and 'Other', along with a 'new question' button. Below the navigation bar, there are filters for 'All questions', 'Open', 'Closed', 'Unanswered', and 'Answered'. The main content area is titled 'Overview Section' and contains a welcome message: 'Welcome on the ELRC Helpdesk!'. It explains that the helpdesk is for EC data users and data providers, and lists several technical topics it covers, such as language resource identification, preparation, processing, and sharing; language resource formatting, encoding, and packaging; and support for basic data processing like cleaning and alignment.



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



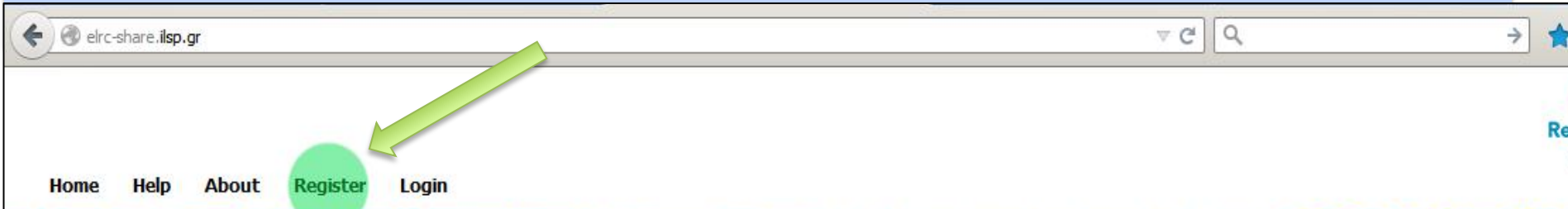
Languages — the heart of
Multilingual Europe



How to Contribute Language Resources (1/7)



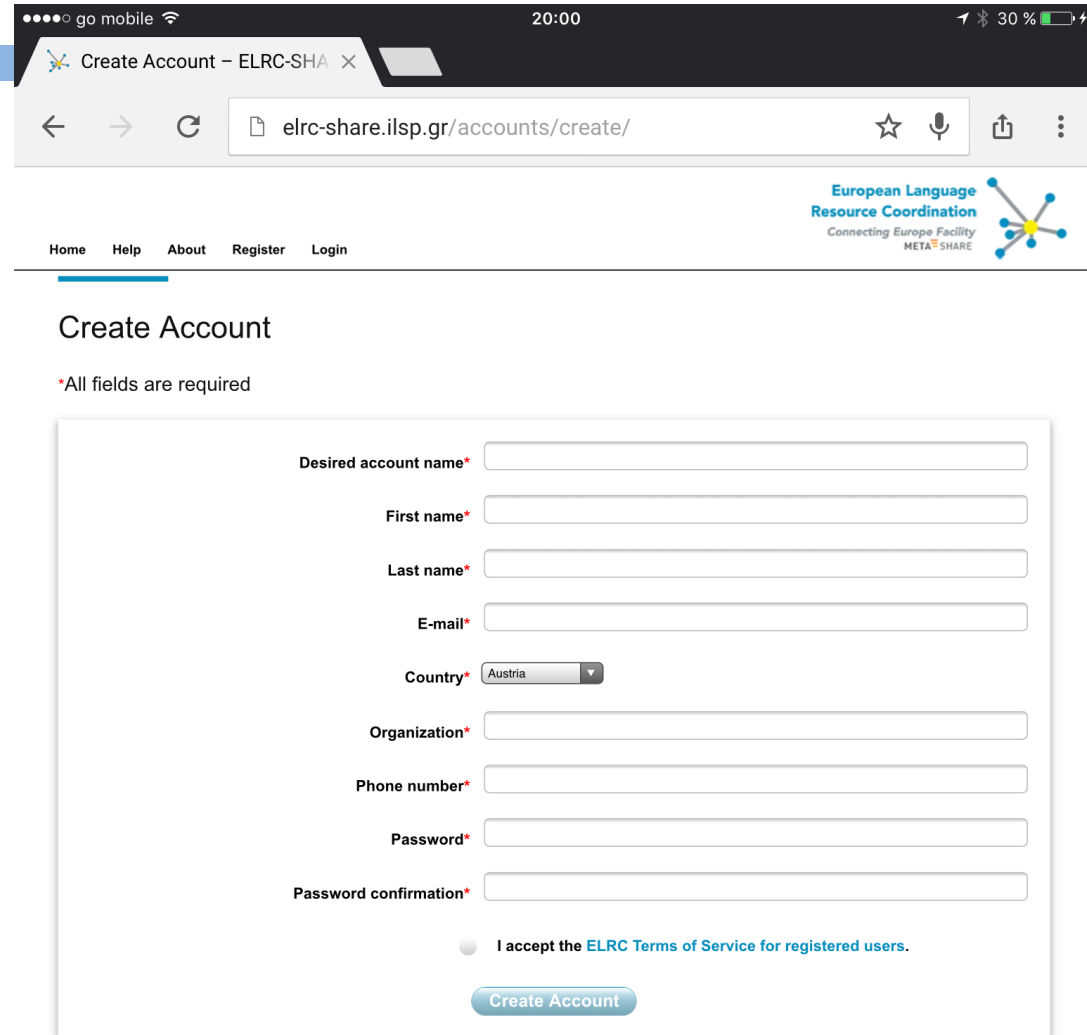
- Go to the ELRC Repository (through the RESOURCE Link)



- Click the *Register* button to get an account



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



go mobile 20:00 30%

Create Account – ELRC-SHA X

elrc-share.ilsp.gr/accounts/create/

European Language Resource Coordination
Connecting Europe Facility
META-SHARE

Home Help About Register Login

Create Account

*All fields are required

Desired account name*

First name*

Last name*

E-mail*

Country* Austria

Organization*

Phone number*

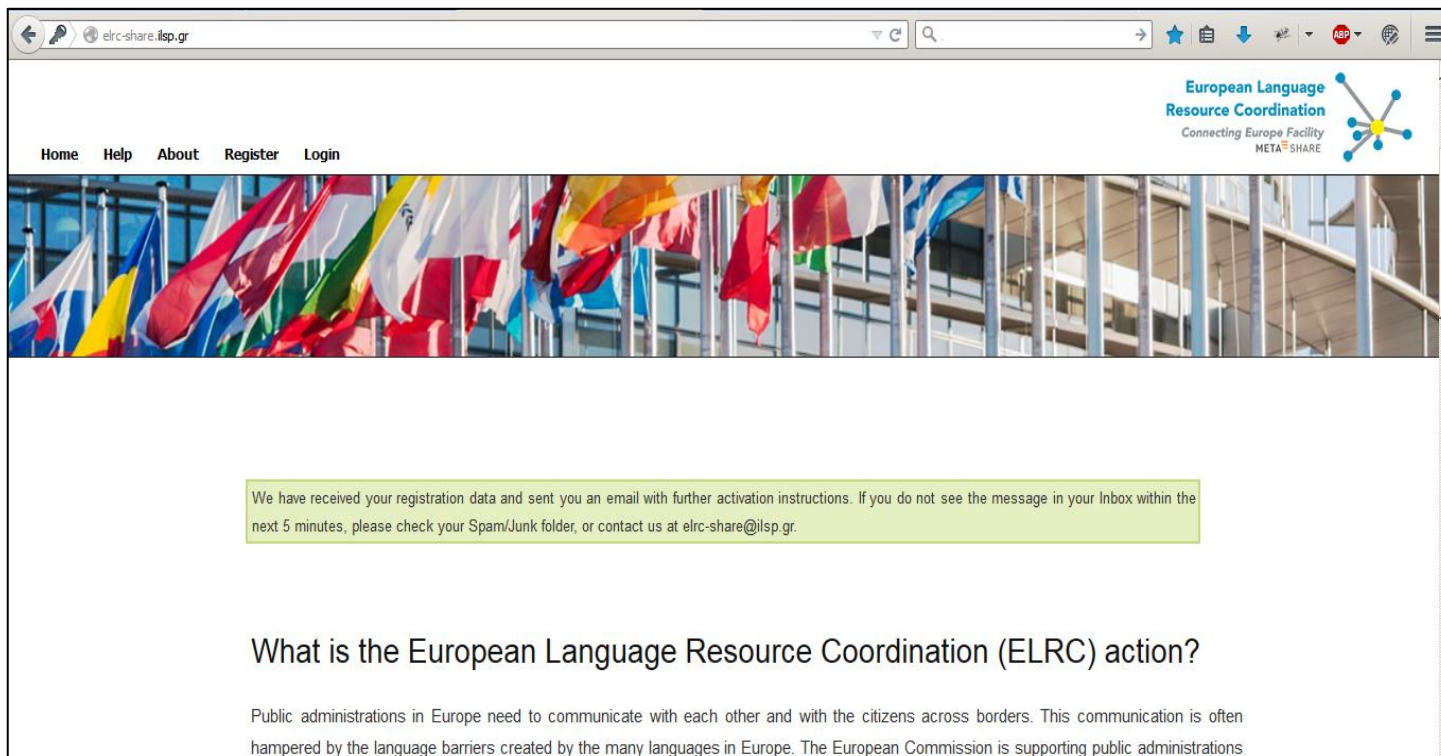
Password*

Password confirmation*

I accept the [ELRC Terms of Service](#) for registered users.

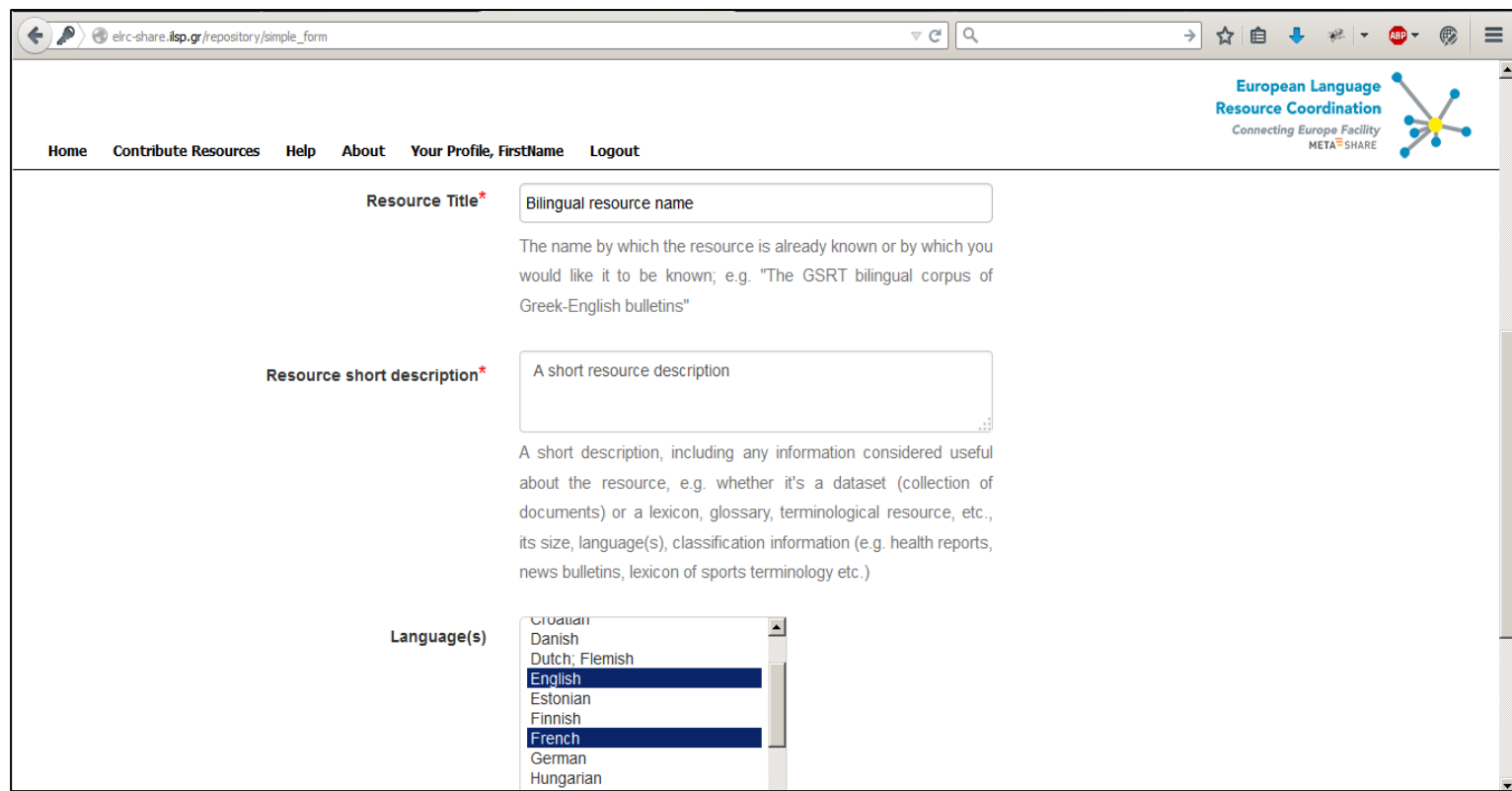
Create Account

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



The screenshot shows a web browser window with the URL `elrc-share.ilsp.gr`. The page header includes the ELRC logo and navigation links: Home, Help, About, Register, Login. Below the header is a banner image of various European national flags. A green message box states: "We have received your registration data and sent you an email with further activation instructions. If you do not see the message in your Inbox within the next 5 minutes, please check your Spam/Junk folder, or contact us at elrc-share@ilsp.gr." Below this, the heading "What is the European Language Resource Coordination (ELRC) action?" is followed by a paragraph: "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

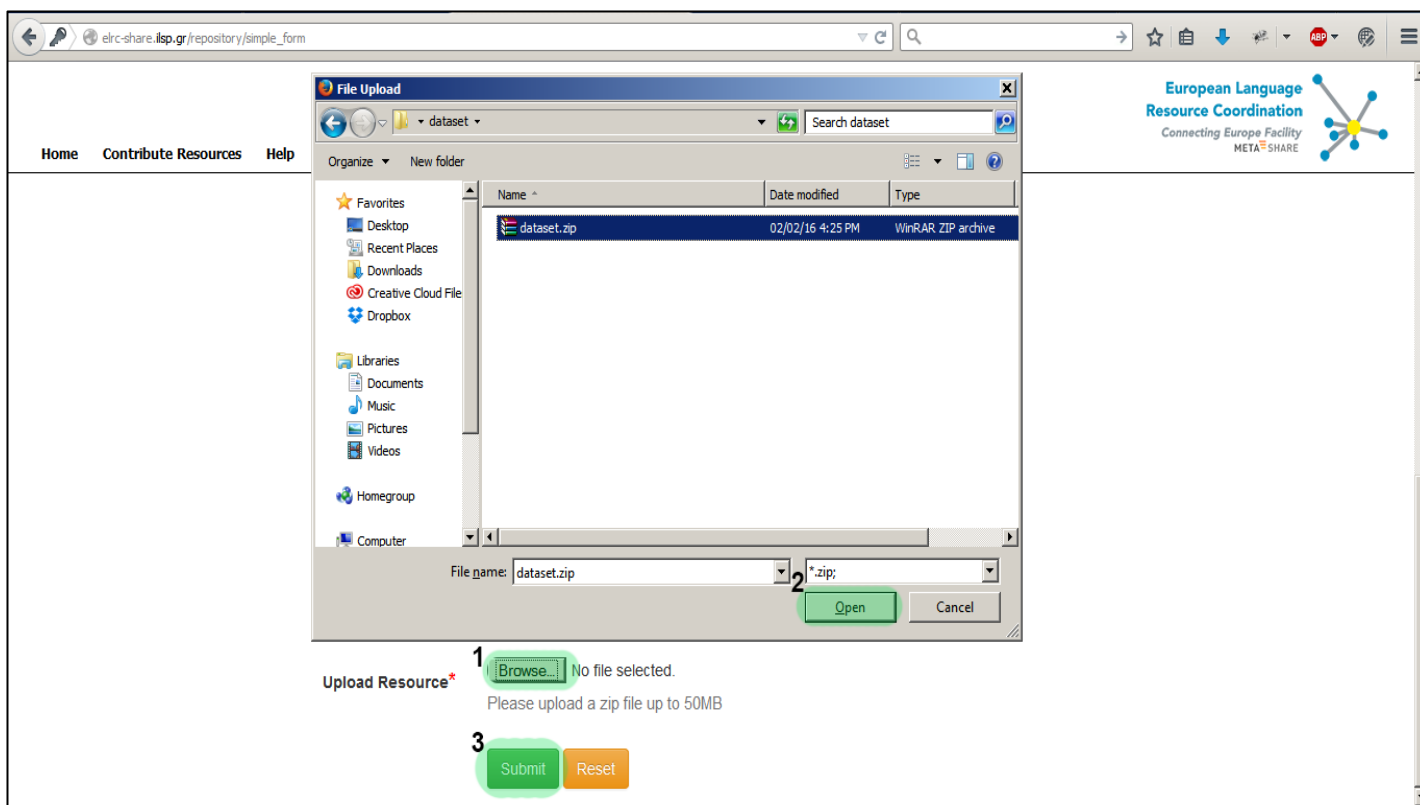
- Fill in the details of the dataset



The screenshot shows a web browser window with the URL `elrc-share.itp.gr/repository/simple_form`. The page header includes the ELRC logo and navigation links: Home, Contribute Resources, Help, About, Your Profile, FirstName, and Logout. The main content area contains three form fields:

- Resource Title***: A text input field containing "Bilingual resource name". Below it is a description: "The name by which the resource is already known or by which you would like it to be known; e.g. 'The GSRT bilingual corpus of Greek-English bulletins'".
- Resource short description***: A text area containing "A short resource description". Below it is a description: "A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size, language(s), classification information (e.g. health reports, news bulletins, lexicon of sports terminology etc.)".
- Language(s)**: A dropdown menu with a scroll bar. The visible options are Croatian, Danish, Dutch; Flemish, English (highlighted), Estonian, Finnish, French (highlighted), German, and Hungarian.

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



1 No file selected.
Please upload a zip file up to 50MB

3

How to Suggest SOURCES of data



Submit
Sources



URL:

Submit URL



*Automatic check if URL already in
database. If not, proceed with submission.*



Source submission form

Source name:

Languages:

Provider name:

....

Contact name:

Email:

Submit Source