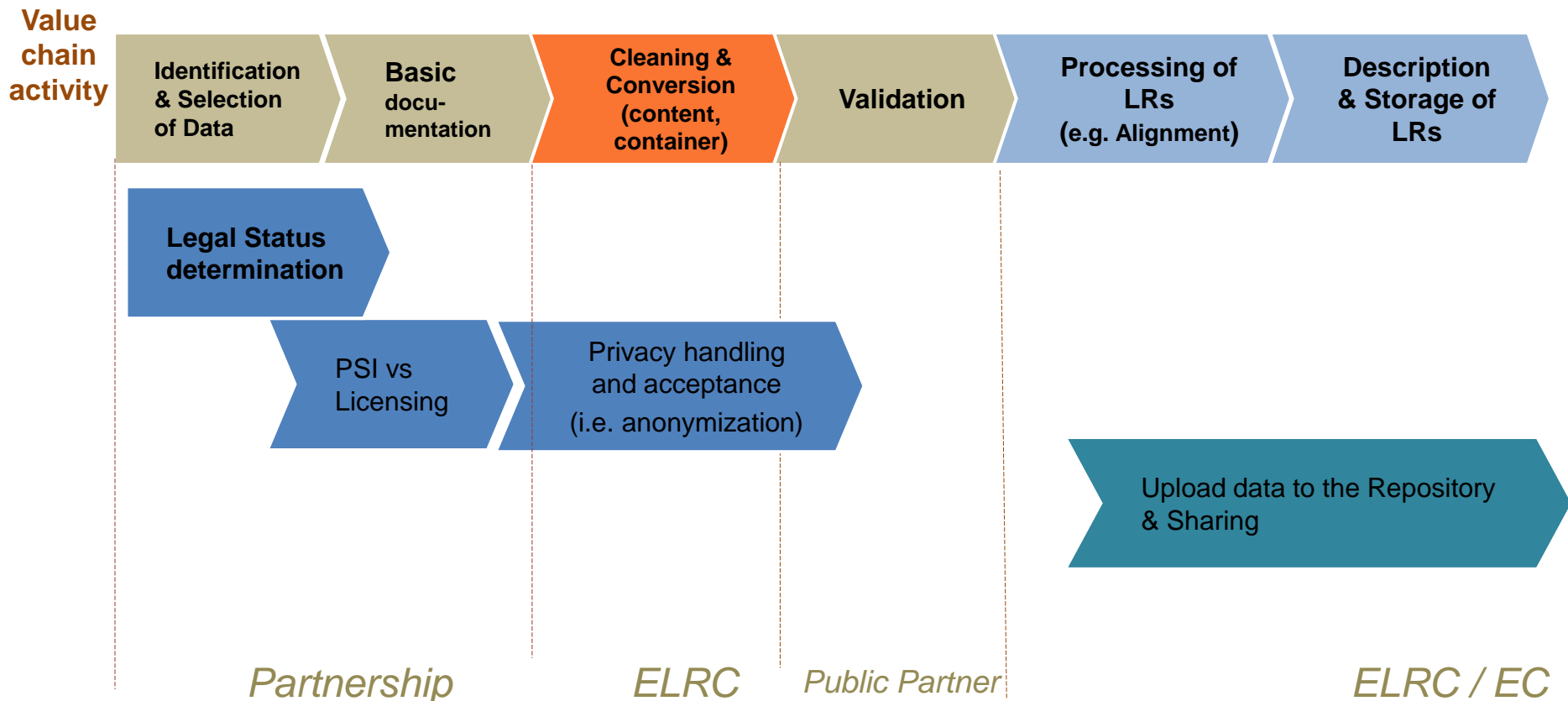


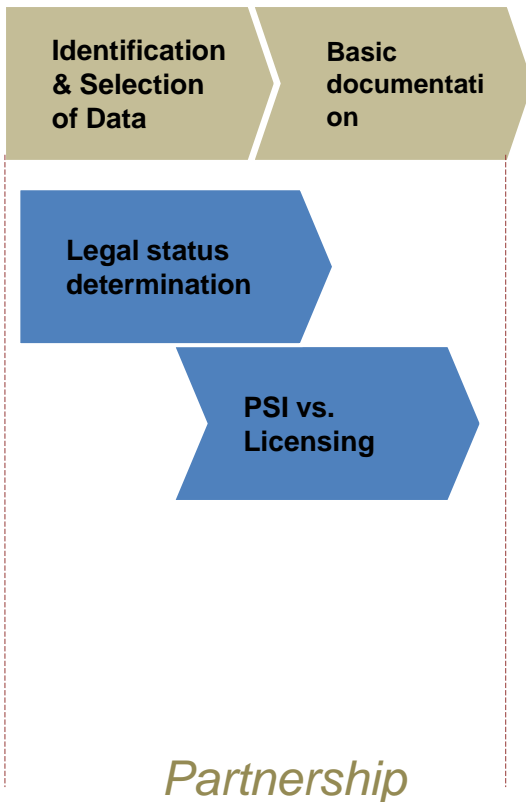
# Sharing Data and Language Resources: Technical Aspects and Best Practices

**Stelios Piperidis**  
**ELRC, ILSP/Athena RC**

# Illustration of data packaging workflow

## Data → LRs (Language Resources)





- Identification of sources
- Identification and selection of data sets (raw data)
- Legal issues
  - Licensing
  - Privacy and ethics management
- Technical issues
  - Choice of Medium and Data formats for the transfer of the “raw” data (preference for the ELRC ad hoc platform)
- Documentation with basic identification elements (Languages, Domains, year, ...)



# Any digital textual data !!





Cleaning &  
Conversion  
(content,  
container)

Privacy handling and  
acceptance  
(i.e. anonymization)

ELRC

### Technical issues

- Cleaning of data format
  - encoding Character sets e.g. UTF8
  - [discarding formatting](#), e.g. bold, italic; graphics, ads, tables, html tags, etc.
  - ...
- File cleaning (e.g. conversion to XML, XLIFF, etc.)
- [Data anonymization](#)



# Formatting example



***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  
οποίοι  
σεις. Ο Διεθνής  
ένα από τα  
ποίος  
στη



- Identify a large source of data on individuals, organizations etc.
  - Use a Named Entity Recognizer (NER) to find and remove private biodata (names, locations, dates, birth information, etc.) and replace with generic placeholders
  - Confirm results meet acceptable requirements
-  Reject data if anonymization is not accurate as required



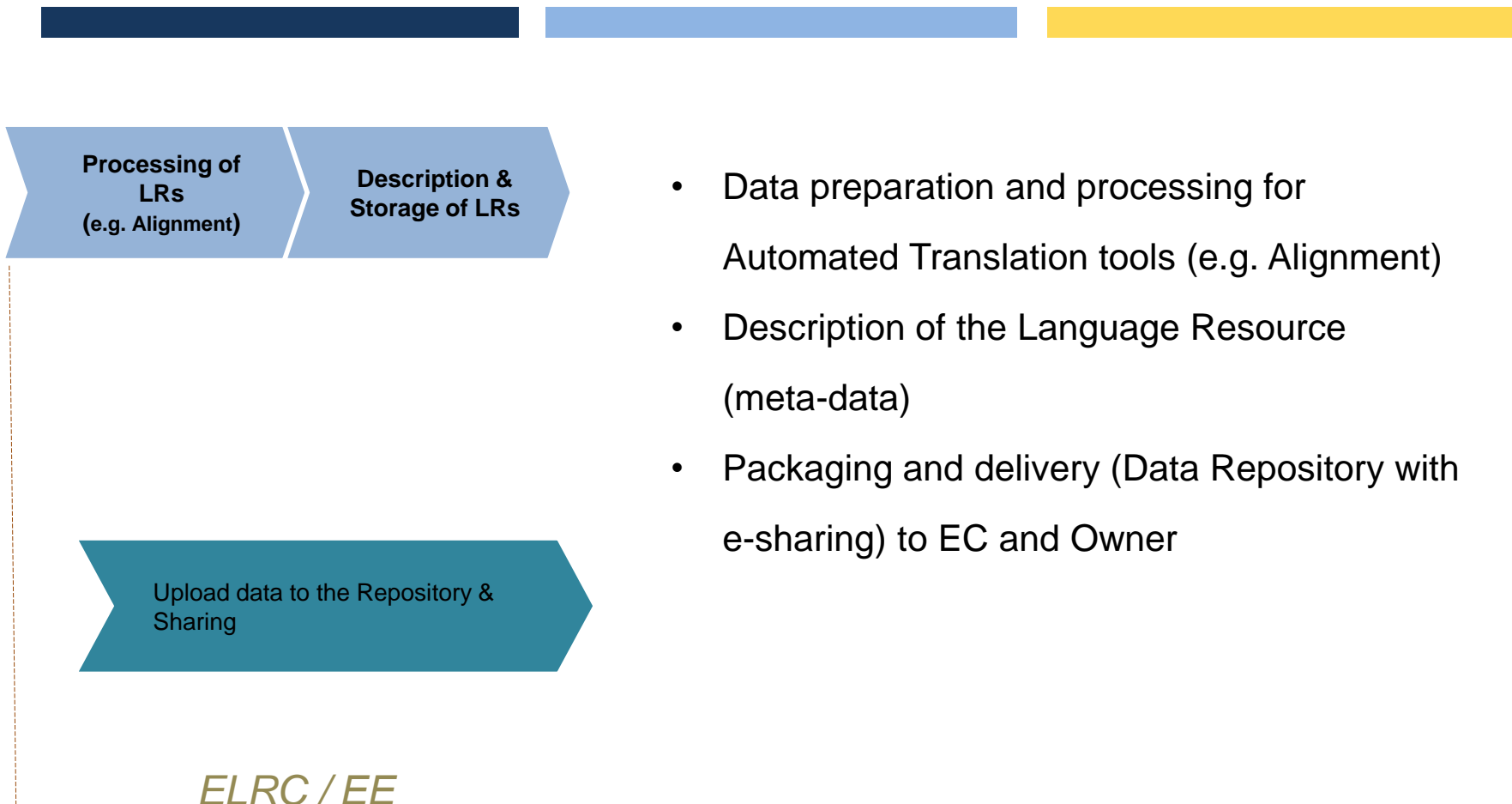
Validation

- Validation and Quality control of the output of the anonymization procedure
- Validation and Quality Control of the output (Language Resource format, content)

→ accept / reject LR

*Public  
partner*







elrc-share.ilsp.gr

European Language  
Resource Coordination  
Connecting Europe Facility  
META-SHARE

Home Help About Register Login

What is the European Language Resource Coordination (ELRC) action?

- 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



- Support for all procedures and technical issues
  - Support services
    - [ELRC portal](#)
    - [technical & legal support helpdesk](#)
    - [forum](#)
    - [repository for sharing LRs](#)



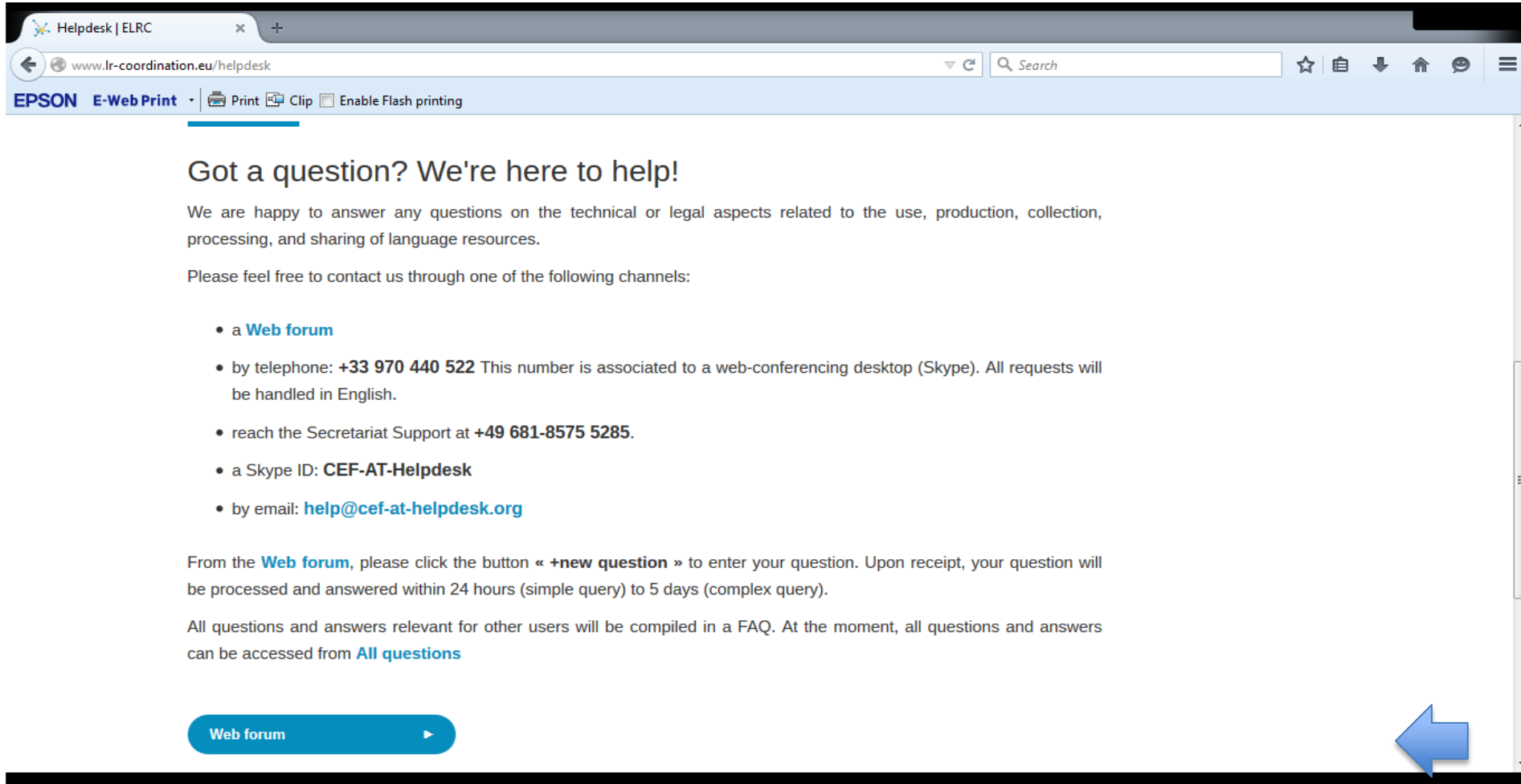
[Home](#) [About](#) [News](#) [Helpdesk](#) [Events](#) [Resources](#) [Anchor Points](#) [Multilingual Europe](#)

European Language  
Resource Coordination



Languages — the heart of  
Multilingual Europe





The screenshot shows a web browser window with the address bar displaying "www.lr-coordination.eu/helpdesk". The page content includes a heading "Got a question? We're here to help!", a paragraph about the helpdesk's purpose, a list of contact channels, and a "Web forum" button. A large blue arrow points to the right in the bottom right corner of the screenshot.

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.

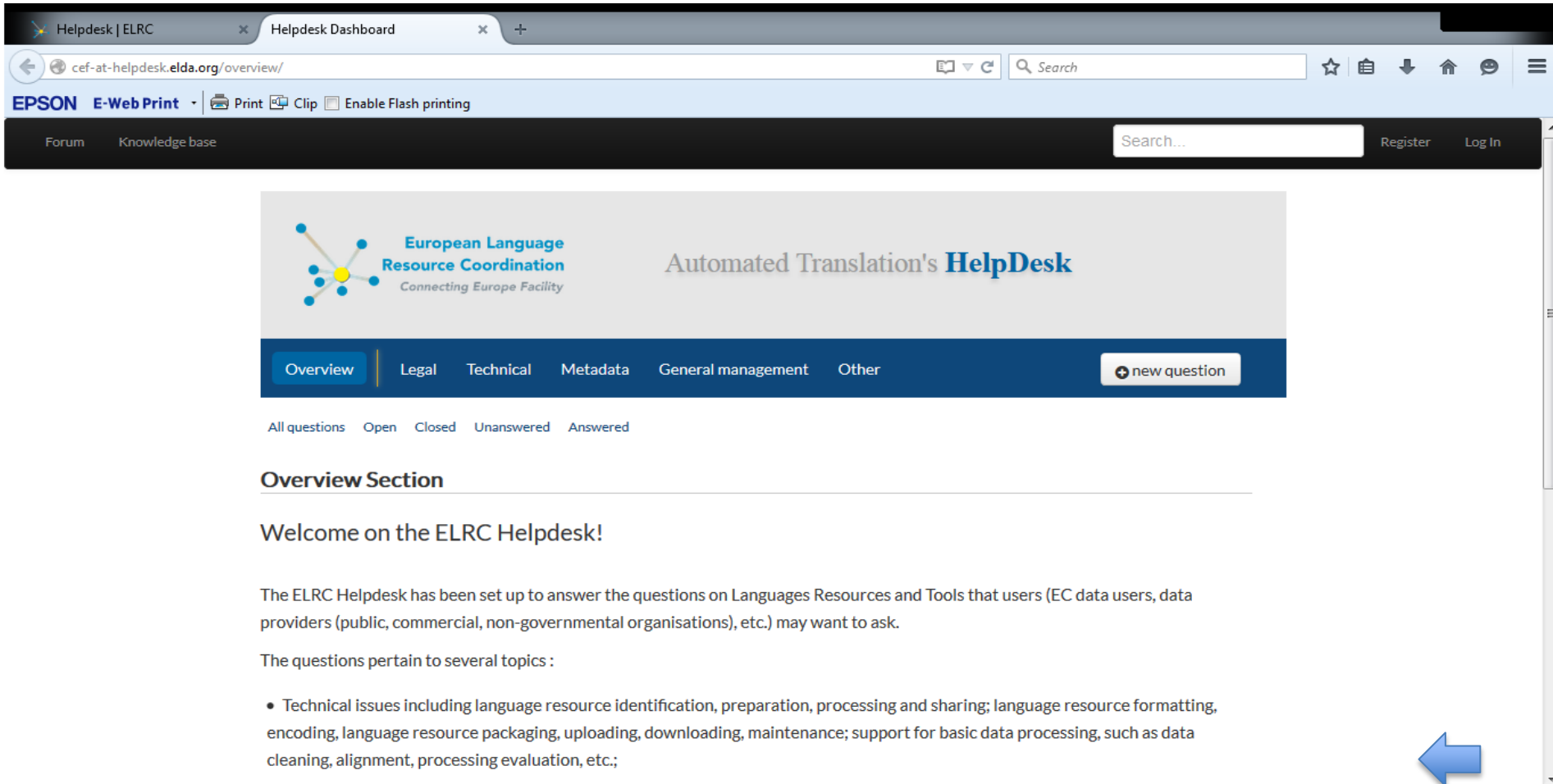
Please feel free to contact us through one of the following channels:

- 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](#)



The screenshot shows a web browser window with two tabs: 'Helpdesk | ELRC' and 'Helpdesk Dashboard'. The address bar shows the URL 'cef-at-helpdesk.elda.org/overview/'. The browser interface includes a search bar, navigation icons, and a toolbar with 'EPSON E-Web Print', 'Print', 'Clip', and 'Enable Flash printing' options. The website header features a dark navigation bar with 'Forum' and 'Knowledge base' links, a search input field, and 'Register' and 'Log In' buttons. The main content area has a light gray header with the ELRC logo and the text 'Automated Translation's HelpDesk'. Below this is a dark blue navigation bar with tabs for 'Overview', 'Legal', 'Technical', 'Metadata', 'General management', and 'Other', along with a '+ new question' button. A filter bar shows 'All questions', 'Open', 'Closed', 'Unanswered', and 'Answered'. The 'Overview Section' is titled 'Welcome on the ELRC Helpdesk!' and contains a welcome message and a list of topics.

Helpdesk | ELRC    Helpdesk Dashboard

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

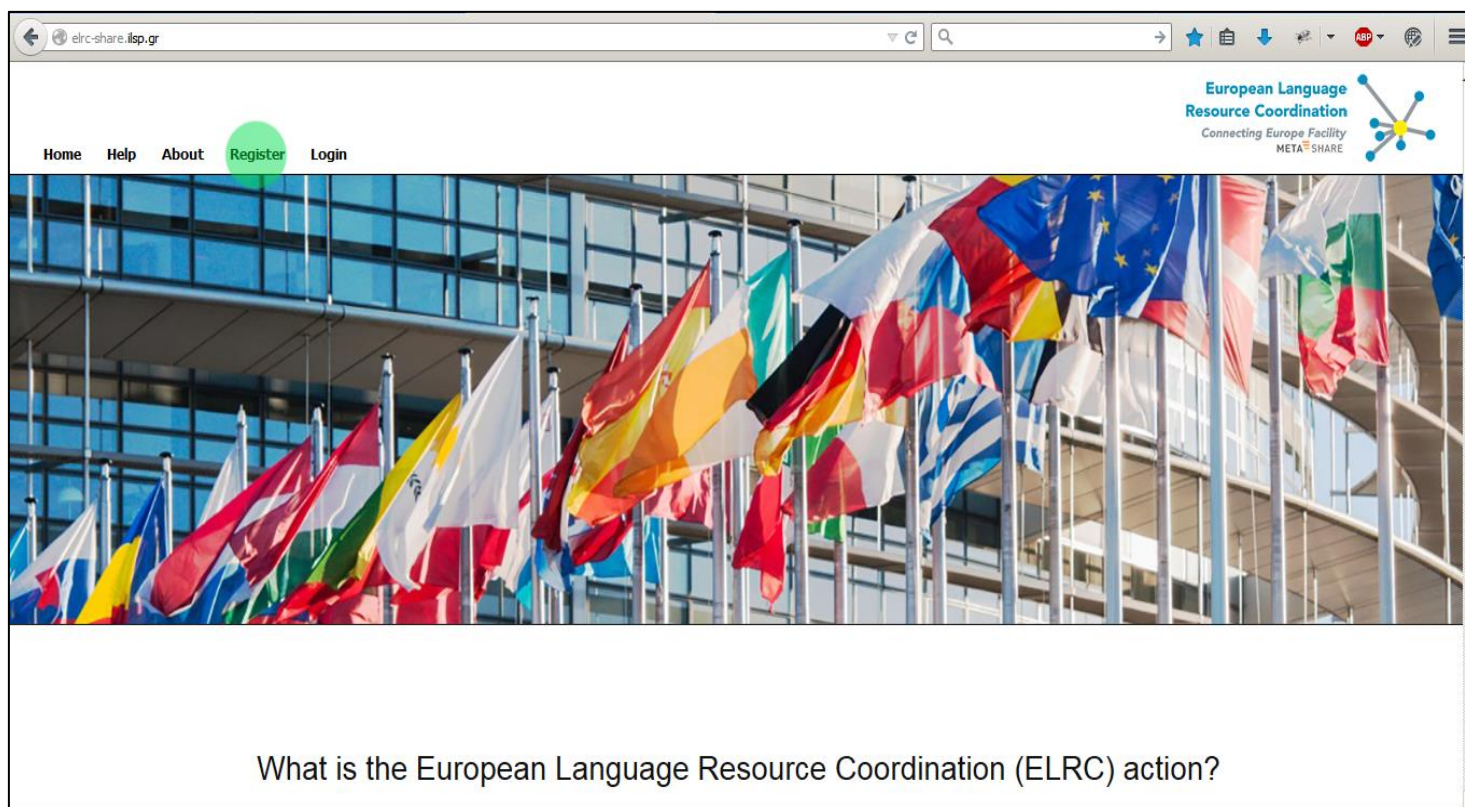
### Welcome on the ELRC Helpdesk!

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.

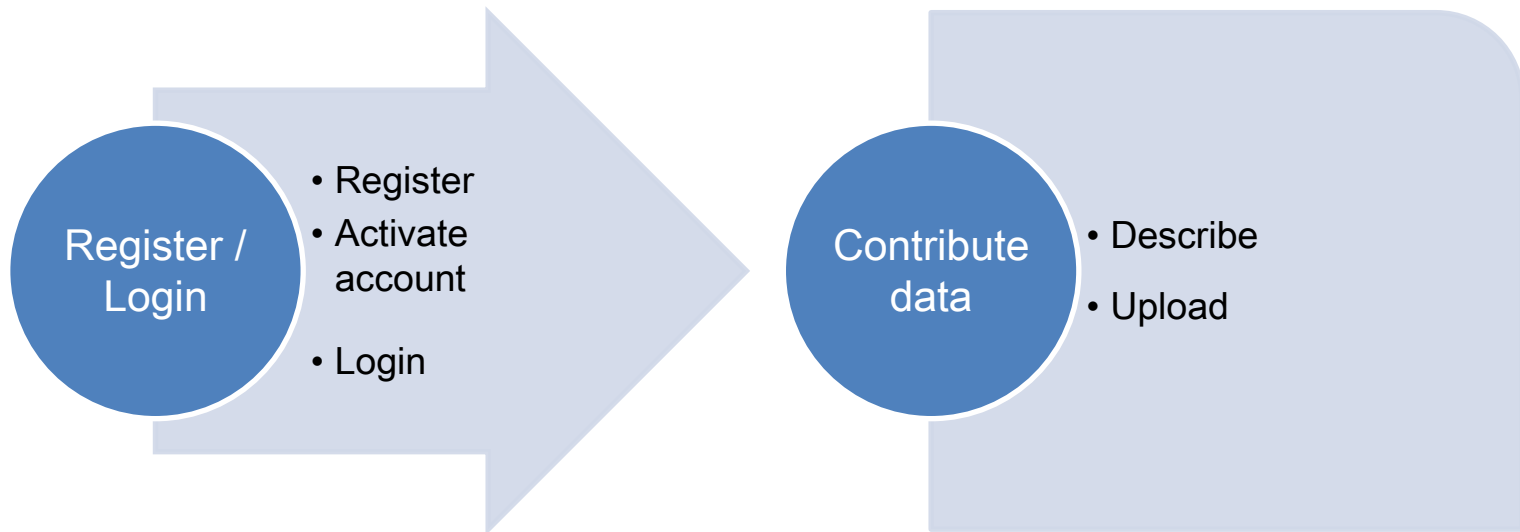
The questions pertain to several topics :

- 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.;

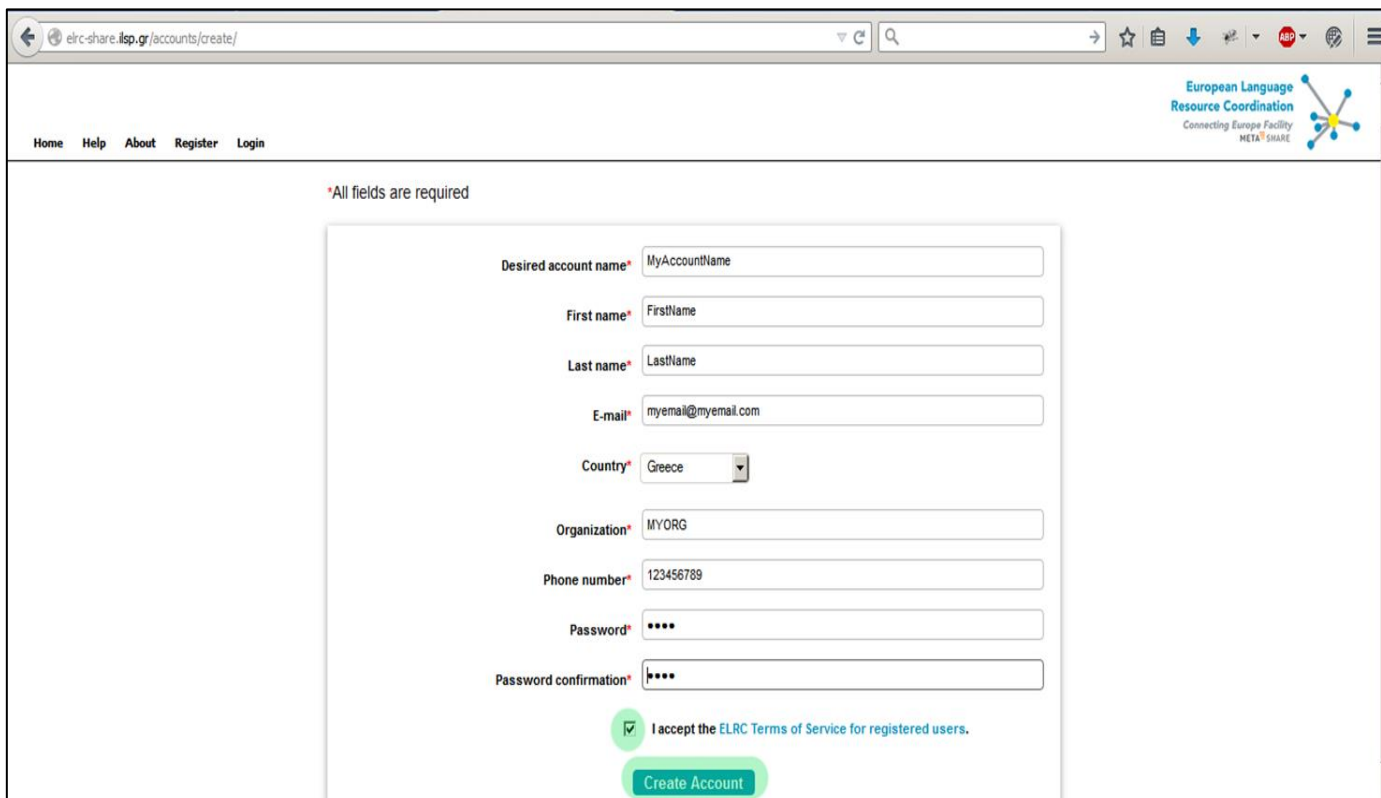
- Go to the ELRC-SHARE Repository: [elrc-share.ilsp.gr](http://elrc-share.ilsp.gr)
- Click the *Register* button







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

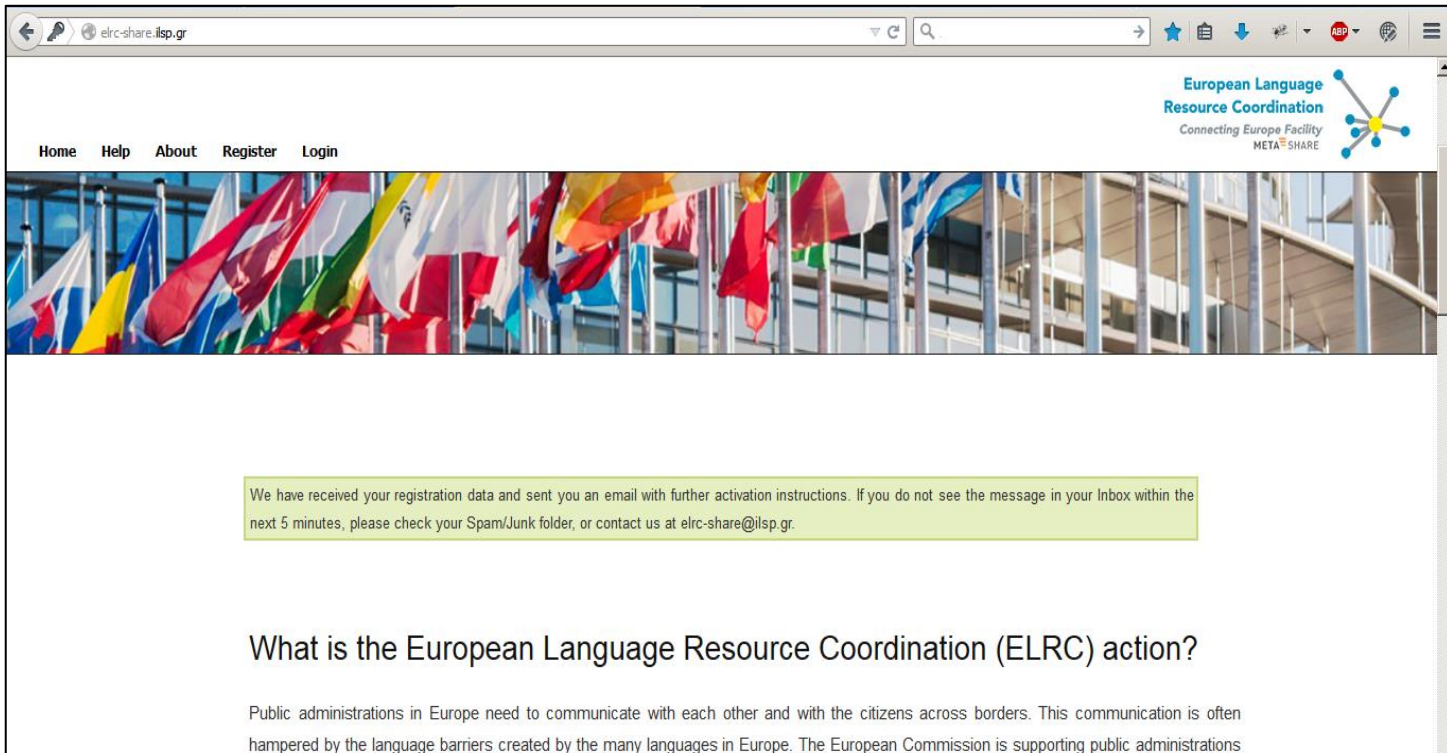


The screenshot shows a web browser window at the URL `elrc-share.isp.gr/accounts/create/`. The page header includes navigation links: Home, Help, About, Register, and Login. The main content area contains a registration form with the following fields:

- Desired account name\* (text input: MyAccountName)
- First name\* (text input: FirstName)
- Last name\* (text input: LastName)
- E-mail\* (text input: myemail@myemail.com)
- Country\* (dropdown menu: Greece)
- Organization\* (text input: MYORG)
- Phone number\* (text input: 123456789)
- Password\* (password input: \*\*\*\*)
- Password confirmation\* (password input: \*\*\*\*)

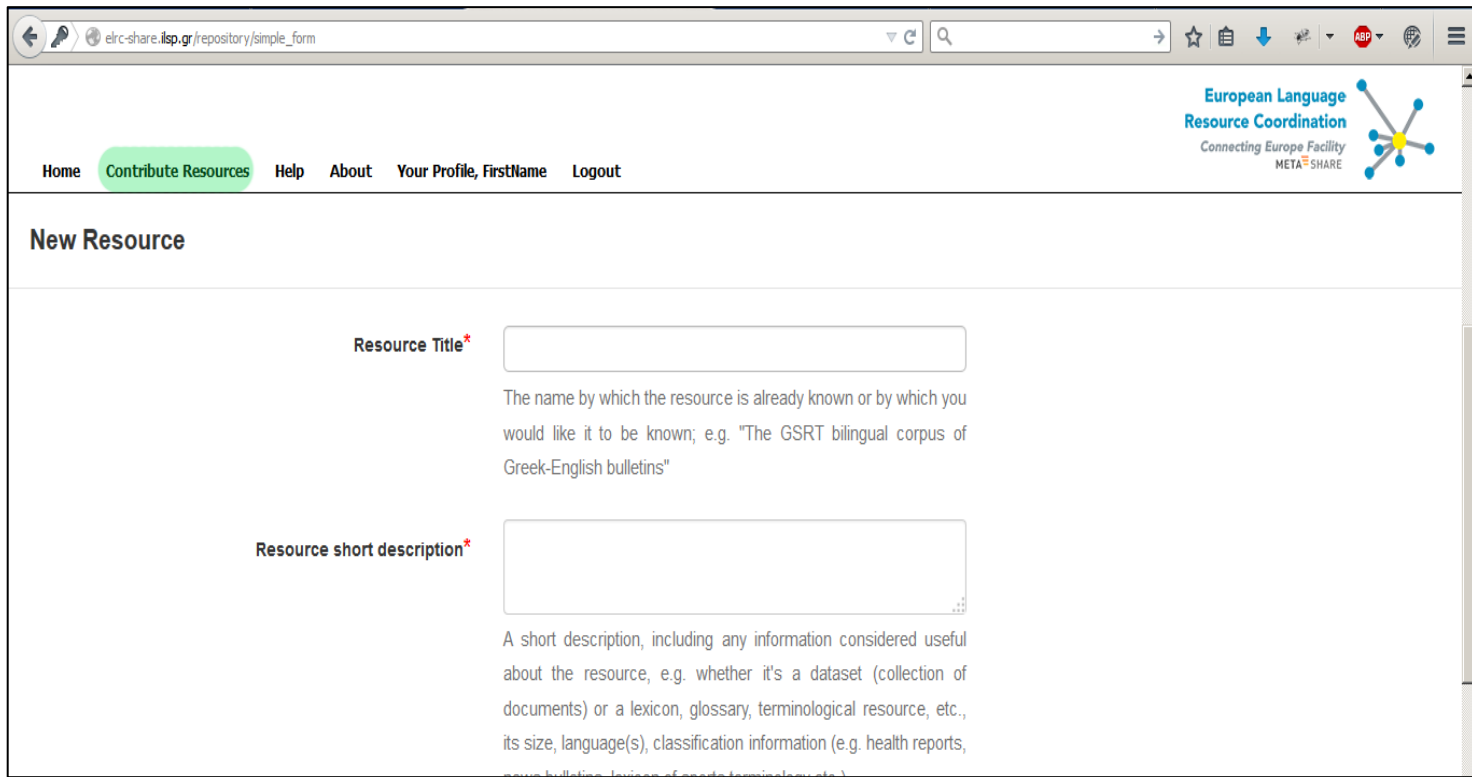
Below the form, there is a checked checkbox labeled "I accept the ELRC Terms of Service for registered users." and a green "Create Account" button.

- 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, and Login. Below the header is a banner image of various European national flags. A green message box in the center of the page reads: "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](mailto:elrc-share@ilsp.gr)." Below the message box, the heading "What is the European Language Resource Coordination (ELRC) action?" is followed by a paragraph of text: "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

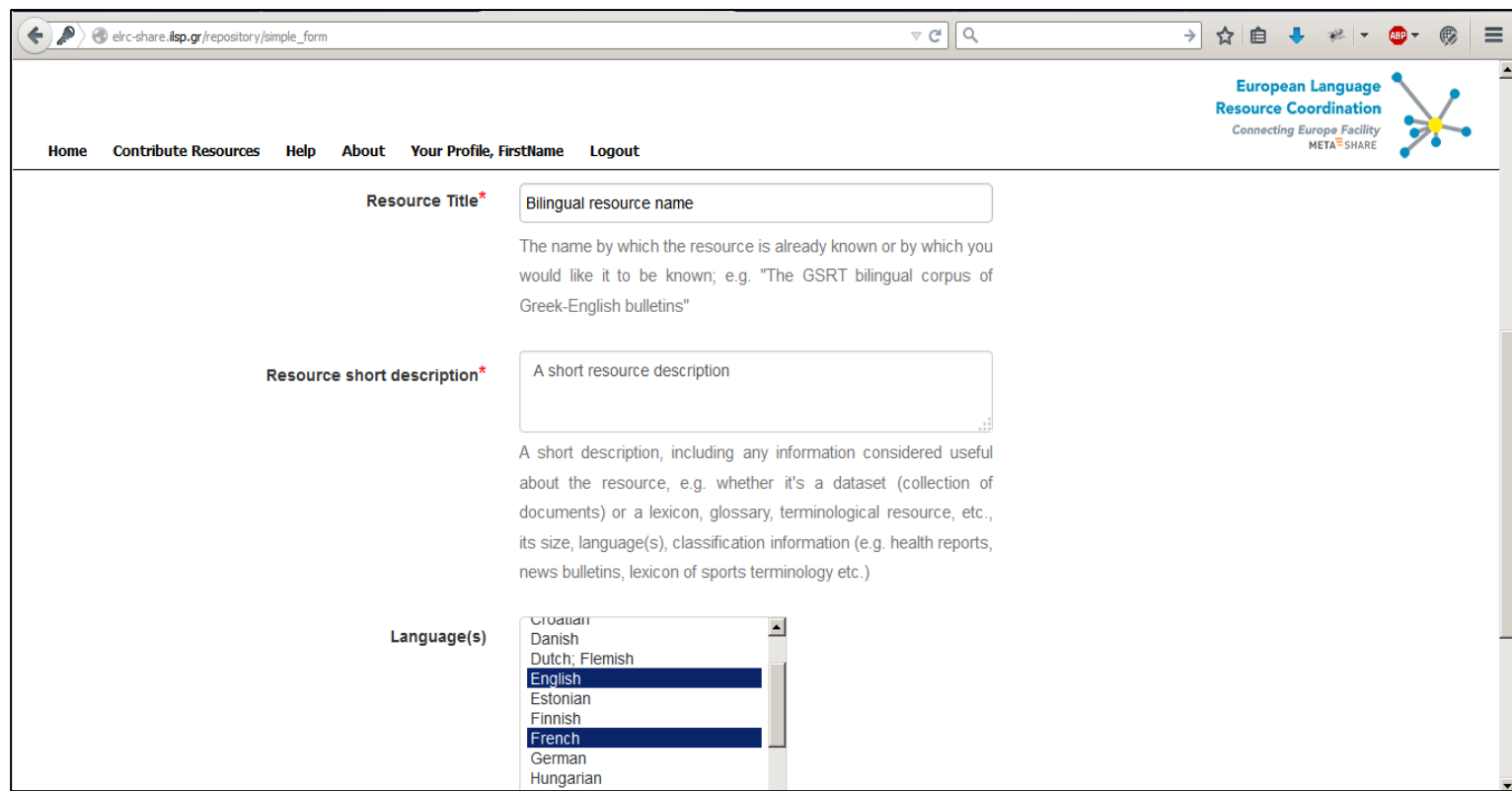
- You get redirected to the data contribution form (or click the Contribute Resources button)



The screenshot shows a web browser window with the URL `elrc-share.jsp.gr/repository/simple_form`. The page header includes the ELRC logo and navigation links: Home, Contribute Resources (highlighted), Help, About, Your Profile, FirstName, and Logout. The main content area is titled "New Resource" and contains two input fields:

- Resource Title\***: A text input field with a placeholder text: "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 with a placeholder text: "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, reviews of books, terminological...)"

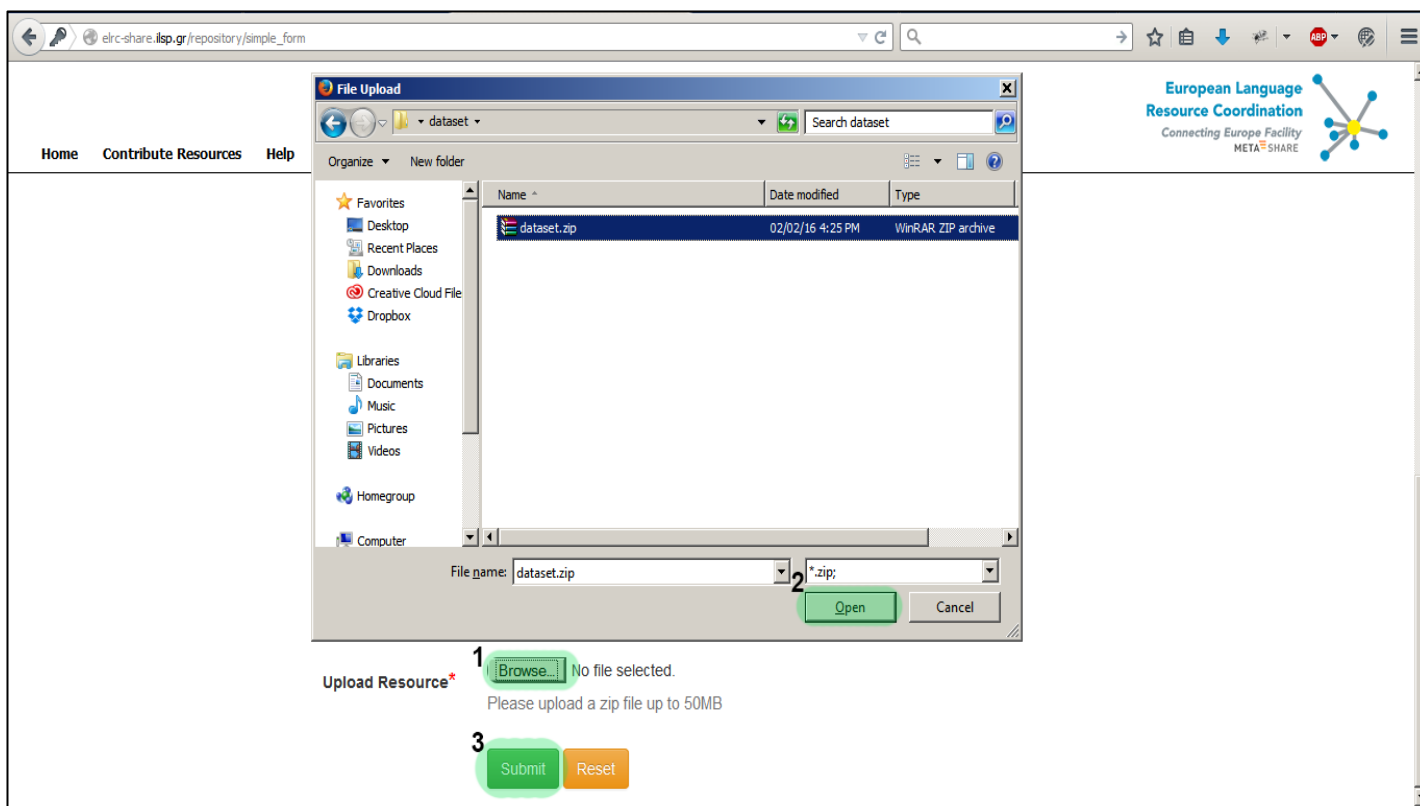
- Fill in the details of the dataset



The screenshot shows a web browser window with the URL `elrc-share.ilsp.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 a form with the following 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 the following options: 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*

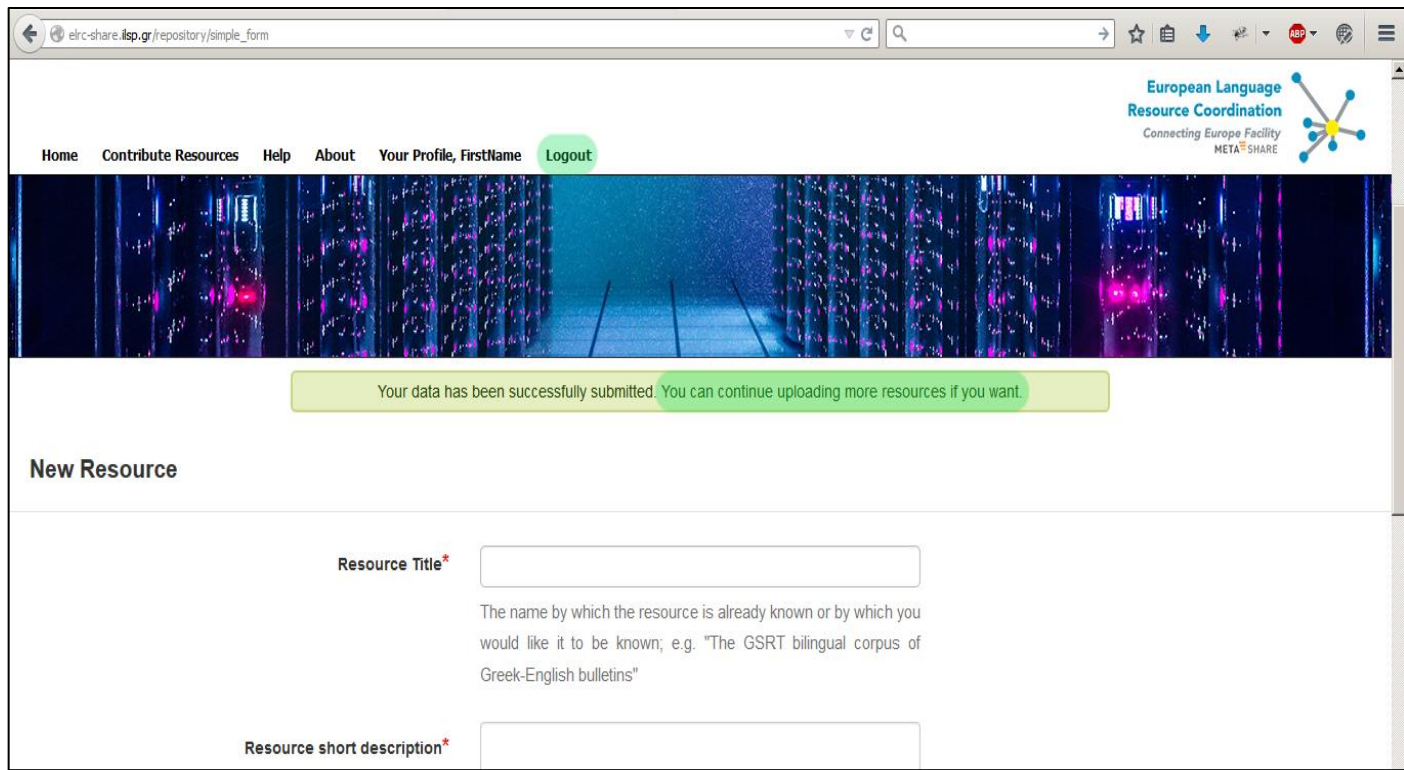


1  No file selected.  
Please upload a zip file up to 50MB

2

3

- Repeat the process if you want to contribute another resource, or log out



The screenshot shows a web browser window at the URL `elrc-share.ilsp.gr/repository/simple_form`. The page header includes the ELRC logo and navigation links: Home, Contribute Resources, Help, About, Your Profile, FirstName, and Logout. A green notification box states: "Your data has been successfully submitted. You can continue uploading more resources if you want." Below this is a "New Resource" section with two input fields: "Resource Title\*" and "Resource short description\*". The "Resource Title\*" field has a placeholder text: "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'".



- 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





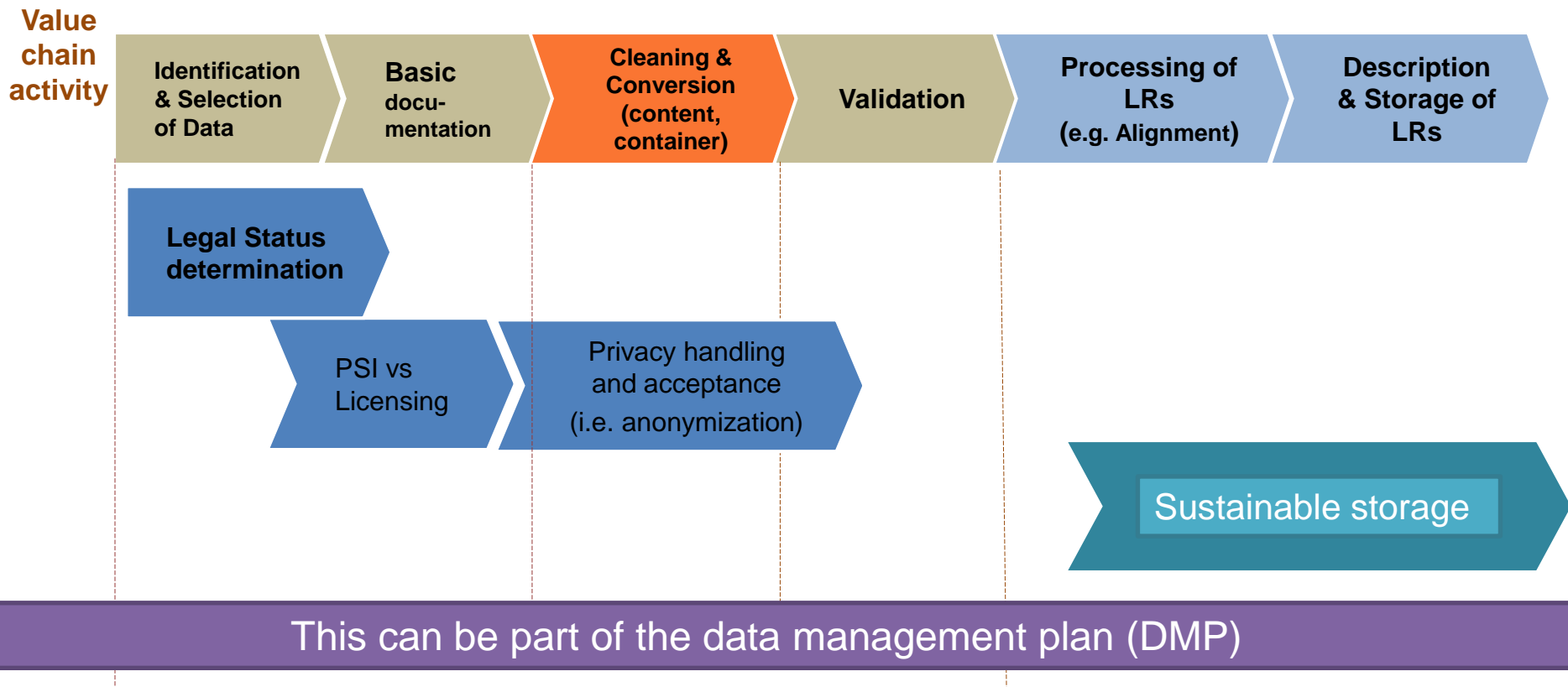
# Best practice for the future: Capitalize on your valuable data

*Best Practice in Data Management*



- Now that I know the value of data, what should my plans be?
- What are the best ways to collect, maintain, archive and re-use my data
- In particular how can I use it for improving MT performances?

# Main phases of data development





- 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 (alignment editors, 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

- See how ELRC can assist you
- There is also the option of national/ European open data portal

