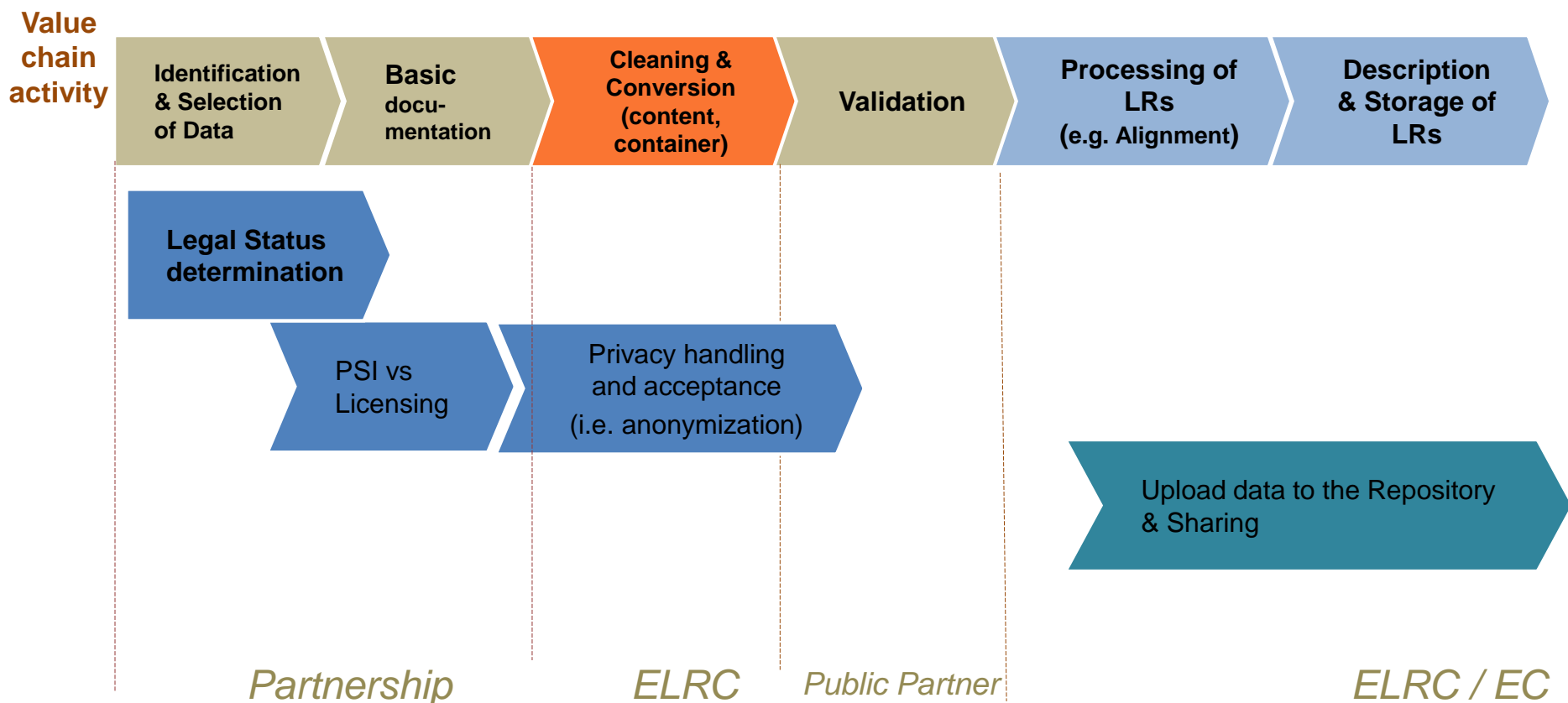


# Sharing Data and Language Resources: Technical Aspects and Best Practices

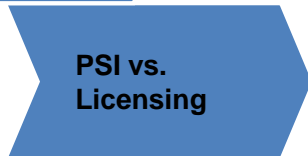
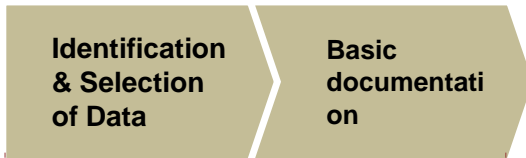
**Stelios Piperidis**  
**ELRC, ILSP/Athena RC**

# Illustration of data packaging workflow

## Data → LR (Language Resources)



# Issues to address (1)



*Partnership*

- 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 οποίων οι οποίοι συμμετέχουν σε 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.



- 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



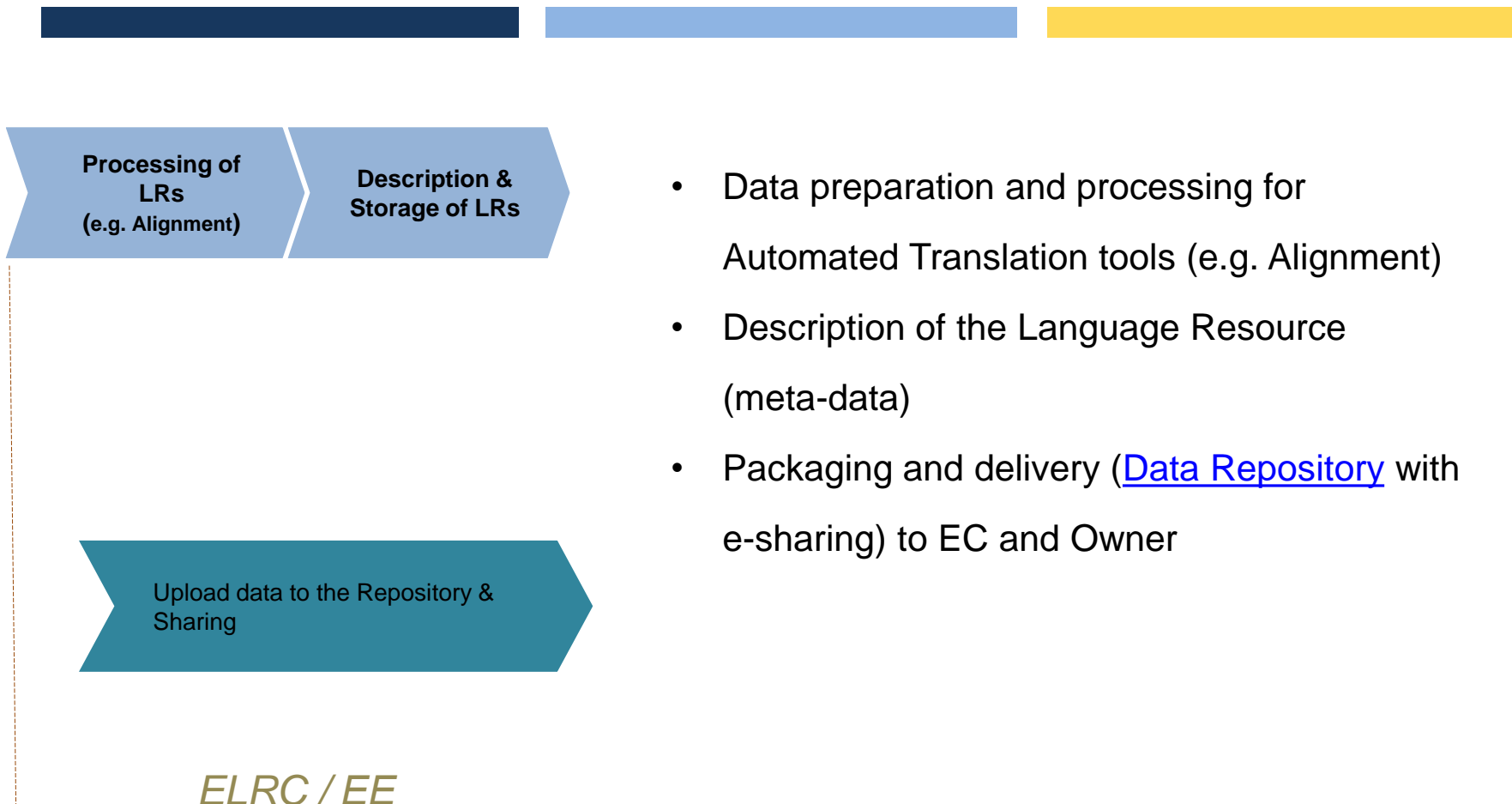
Validation

- 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*

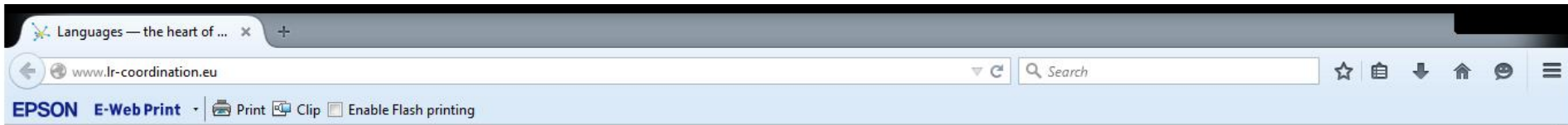




- Identification of sources
- Identification and selection of data sets (raw data)
  - Data can be obtained from the visible sources (e.g. harvested from web)
  - Data can be handed over by the public sector players
  - Public sector players can boost the identification of visible sources
- Processing 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](#)
    - [repository for sharing LRs](#)
    - [forum](#)



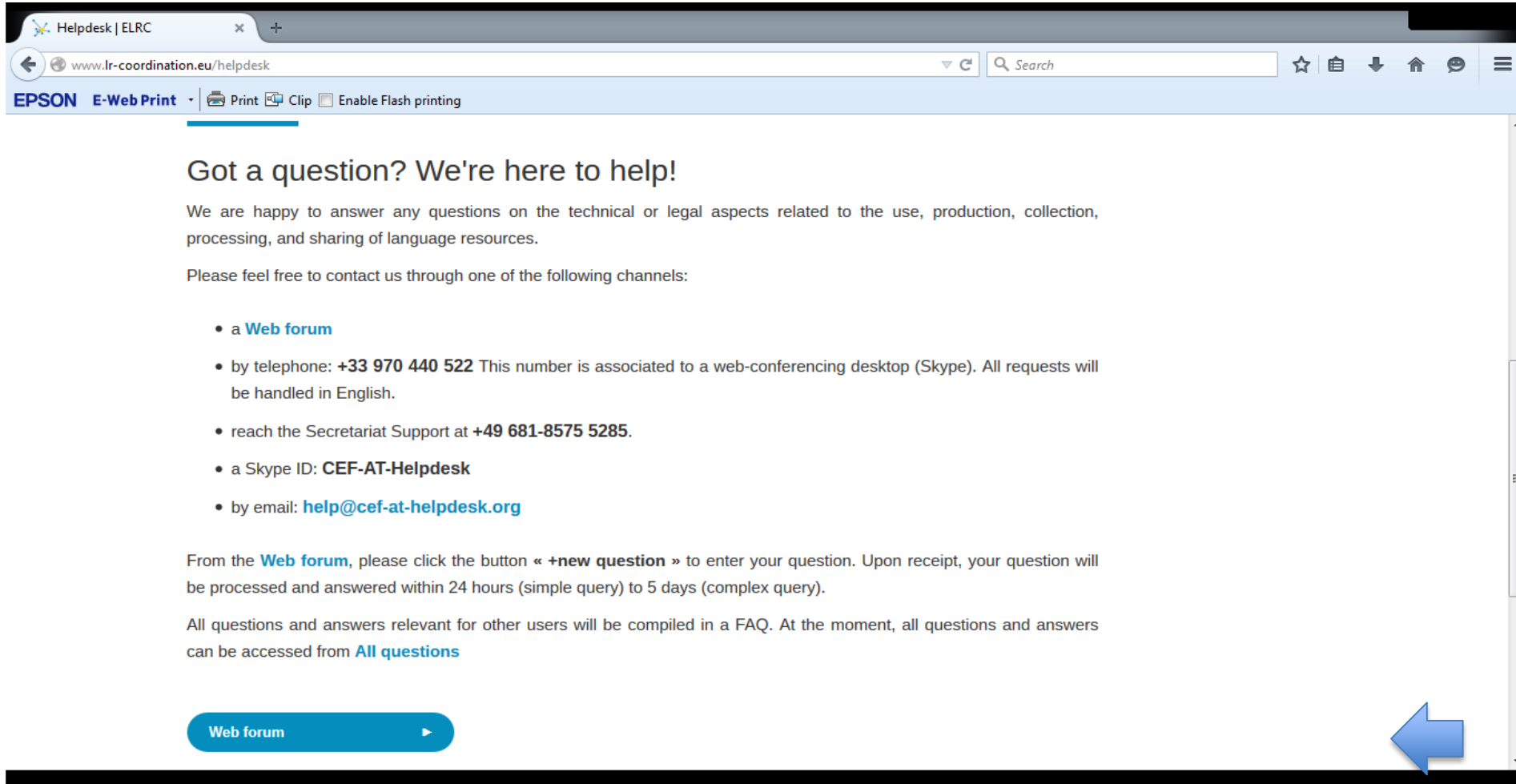
[Home](#) [About](#) [News](#) [Helpdesk](#) [Events](#) [Resources](#) [Anchor Points](#) [Multilingual Europe](#)

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 blue arrow points to 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.

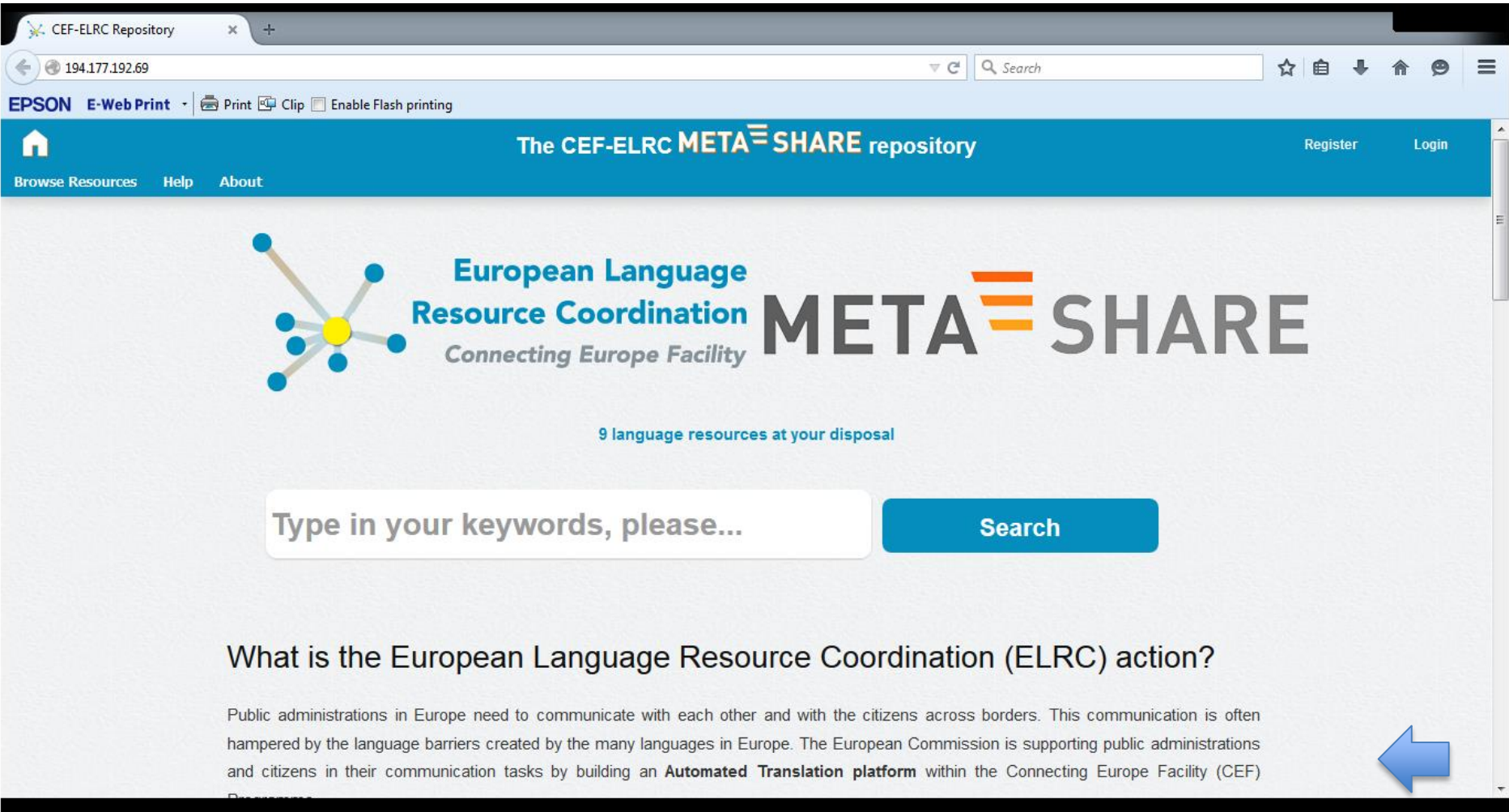
Please feel free to contact us through one of the following channels:

- 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 displaying the CEF-ELRC Repository website. The browser's address bar shows the URL 194.177.192.69. The website's header features the text "The CEF-ELRC META SHARE repository" and navigation links for "Register" and "Login". Below the header, there are links for "Browse Resources", "Help", and "About". The main content area displays the European Language Resource Coordination logo and the text "META SHARE". A search bar with the placeholder text "Type in your keywords, please..." and a "Search" button is present. Below the search bar, the text "9 language resources at your disposal" is displayed. The bottom section of the page contains the heading "What is the European Language Resource Coordination (ELRC) action?" followed by a paragraph of text. A blue arrow points to the right in the bottom right corner of the screenshot.

CEF-ELRC Repository

194.177.192.69

EPSON E-Web Print Print Clip Enable Flash printing

The CEF-ELRC META SHARE repository

Register Login

Browse Resources Help About

European Language Resource Coordination Connecting Europe Facility

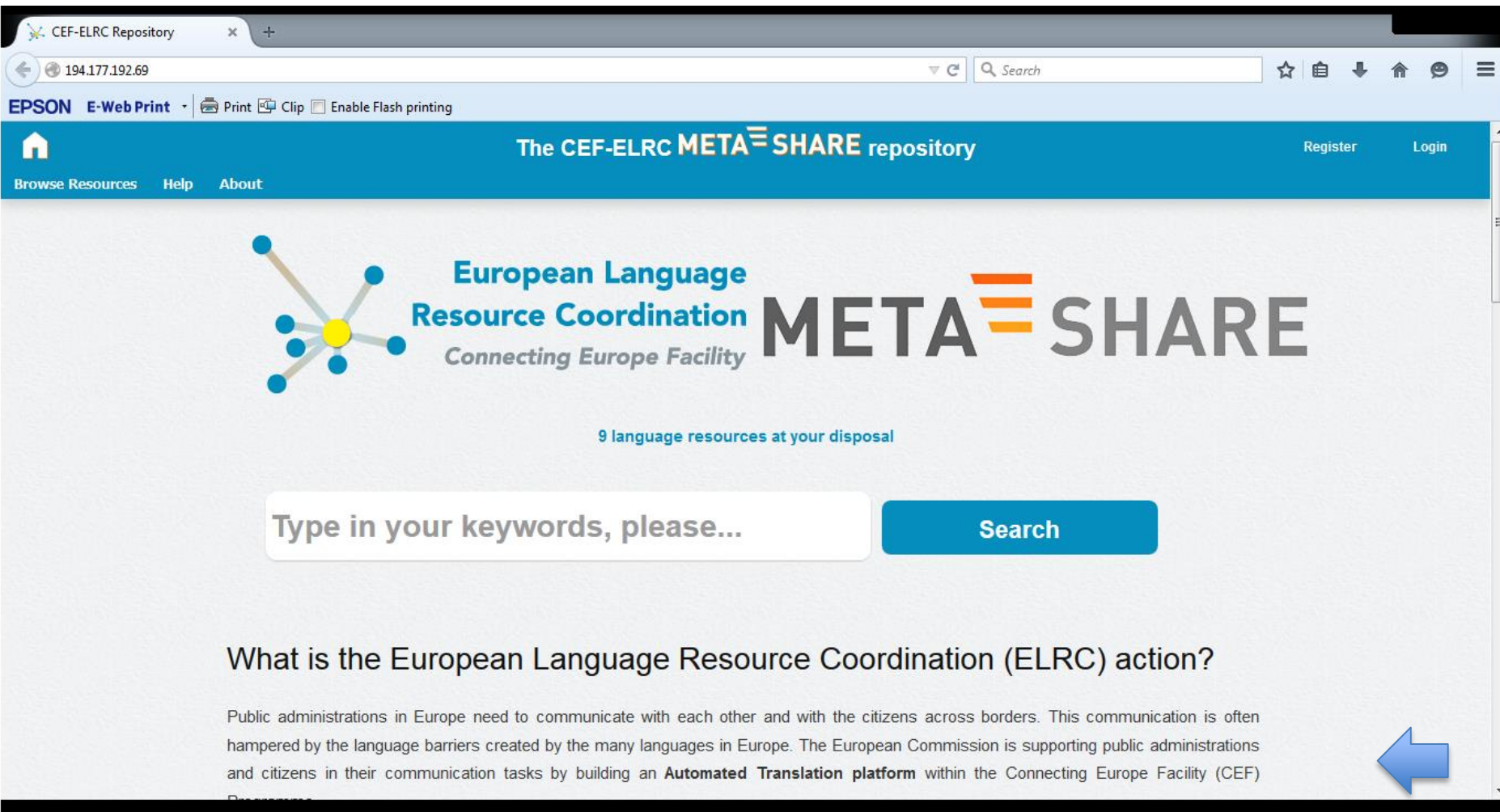
META SHARE

9 language resources at your disposal

Type in your keywords, please... Search

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

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 and citizens in their communication tasks by building an **Automated Translation platform** within the Connecting Europe Facility (CEF)



The screenshot shows a web browser window displaying the CEF-ELRC Repository website. The browser's address bar shows the URL 194.177.192.69. The website's header features the text "The CEF-ELRC META SHARE repository" and navigation links for "Register" and "Login". Below the header, there are links for "Browse Resources", "Help", and "About". The main content area displays the European Language Resource Coordination logo and the text "Connecting Europe Facility". To the right of the logo, the words "META SHARE" are prominently displayed. Below this, it states "9 language resources at your disposal". A search bar with the placeholder text "Type in your keywords, please..." and a blue "Search" button is visible. At the bottom of the page, there is a section titled "What is the European Language Resource Coordination (ELRC) action?" followed by a paragraph of text. A blue arrow points to the right in the bottom right corner of the screenshot.

CEF-ELRC Repository

194.177.192.69

EPSON E-Web Print Print Clip Enable Flash printing

The CEF-ELRC META SHARE repository

Register Login

Browse Resources Help About

European Language Resource Coordination Connecting Europe Facility

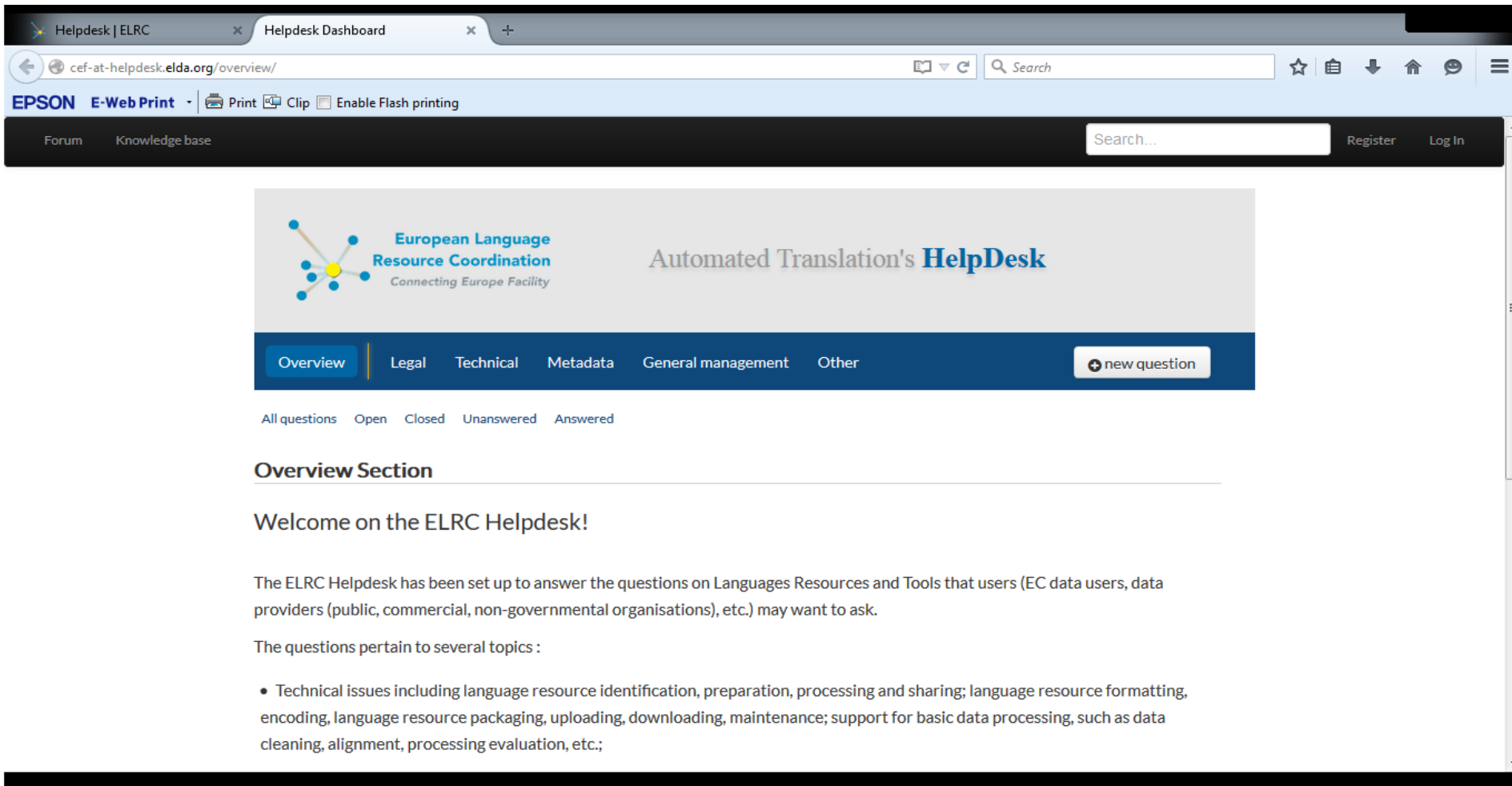
META SHARE

9 language resources at your disposal

Type in your keywords, please... Search

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

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 and citizens in their communication tasks by building an **Automated Translation platform** within the Connecting Europe Facility (CEF)



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 header includes 'EPSON E-Web Print', 'Print', 'Clip', and 'Enable Flash printing' options. A navigation bar contains 'Forum', 'Knowledge base', a search box, and 'Register' and 'Log In' links. The main content area features the ELRC logo and the title 'Automated Translation's HelpDesk'. Below this is a navigation menu with 'Overview', 'Legal', 'Technical', 'Metadata', 'General management', and 'Other' tabs, along with a '+ new question' button. A filter section shows 'All questions', 'Open', 'Closed', 'Unanswered', and 'Answered' options. The 'Overview Section' is titled 'Welcome on the ELRC Helpdesk!' and contains a paragraph explaining the helpdesk's purpose and a list of topics it covers.

Helpdesk | ELRC x Helpdesk Dashboard x +

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.;





- 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



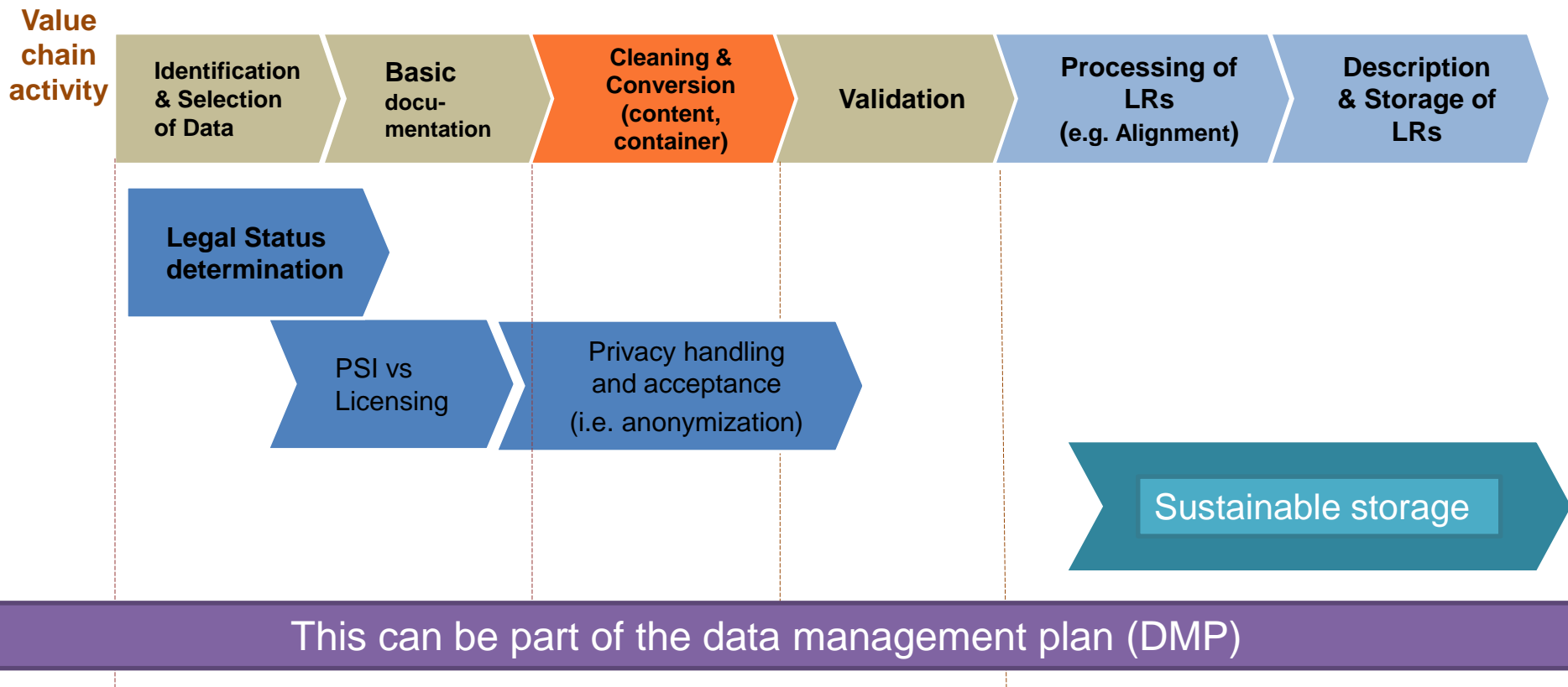
# Best practice for the future: Capitalize on your valuable data

*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

