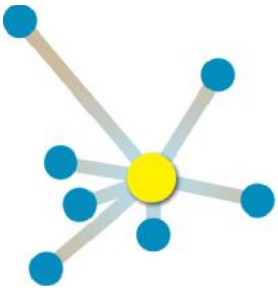
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.



European Language Resource Coordination
Connecting Europe Facility

Deliverable Task 2

Resource Collection Guidelines for NAPs

Author(s): Khalid Choukri (ELDA)
Miltos Deligiannis (ILSP)
Penny Labropoulou (ILSP)
Andrea Lösch (DFKI)
Valérie Mapelli (ELDA)
Stelios Piperidis (ILSP)

Dissemination Level: Public

Version No.: <V1.2>

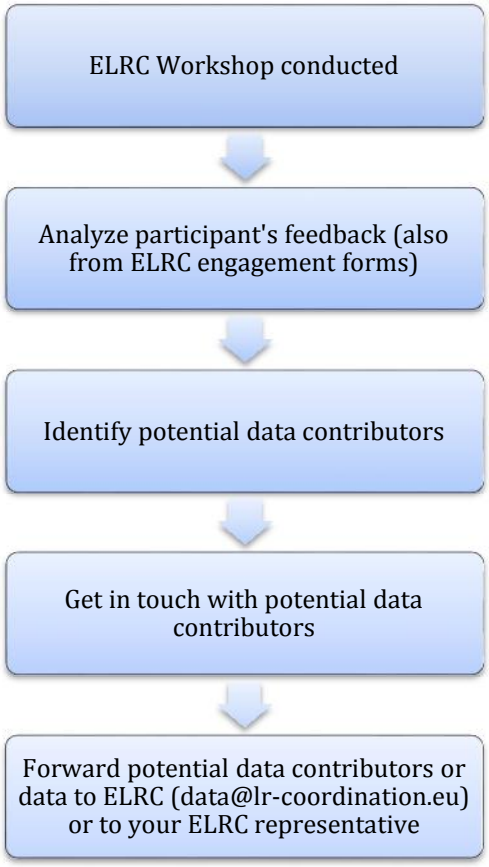
Date: 2016-05-20

Contents

1 Data Collection Process

1.1 How to find data contributors?

In most countries, the ELRC workshops are the kick-off for the data collection process. To the workshop, you should have invited potential data holders from your national public sector administration and public bodies. As such, the ideal process for turning your workshop into the successful data collection involves a direct follow up with your workshop participants:



Timeline: 2 weeks after the workshop

Timeline: 3-4 weeks after the workshop

Timeline: 4 weeks after the workshop

Any time!!!

At any time, your ELRC representative is here to help. The list of the ELRC representative responsible for your country is available in Annex 1.

Important note: In many cases it might be a good idea to personally get in touch with potential data contributors. So if an email remains unanswered, please try to get in touch with the particular person directly.

For your information, example lists of potential data holders are provided in Annex 2.

1.2 How to provide data to ELRC?

Data can principally be donated through the ELRC channel and in very specific cases (for confidentiality, privacy reasons, etc.) to the European Commission.

ELRC offers different options from which data contributors can choose:

- Send data to ELRC via email: data@lr-coordination.eu
- Upload data directly into the ELRC repository through a simple web form: See access point on the ELRC website (http://www.lr-coordination.eu/resources); shortly after the submission, an ELRC member will contact the contributor to obtain further information about the data in order to enrich its description (see Annex 1 for ELRC representatives for data collection in each country).

For those providers wanting to deliver their data directly to the **European Commission**, they can directly get in touch with the representative of their DGT: Szymon KLOCEK, Machine Translation Quality Officer & Machine Translation Data Administrator, Directorate-General for Translation, Unit R.3 – IT, ARIA 03/B080 Luxemburg (szymon.klocek@ec.europa.eu, +352 4301-33543)). If contact is established by email, the email should be sent to szymon.klocek@ec.europa.eu and cc to CEF-AT@ec.europa.eu. Together with the EC, they can then define how best to deliver the data, e.g. whether to send the data by email or whether to upload it to a secure ftp server provided by the EC's DGT etc. The ELRC consortium will be informed about the data provided by the EC.

1.3 What kind of data does ELRC actually look for?

ELRC is looking for language resources from national public sector administrations and public bodies in all EU member states, Norway and Iceland. As such, **any language resource produced or relevant to public sector administrations and public bodies** are useful and welcome by ELRC. The data can take whatever shape. A non-exhaustive list of typical datasets that is sought by ELRC includes:

- Translation memories: linguistic databases that capture translations made by humans. They can be used to facilitate future translations tasks but also for training automated translation systems
- Corpora: ideally multilingual corpora, comparable, aligned, parallel documents, but also, if relevant, monolingual corpora
- Lexica: monolingual and multilingual lists of words, multi-words, sentences, etc. in general or specific subject fields
- Terminological resources: structured sets of concepts, with associated linguistic information in a specific subject field

In general, **language resources** should cover as many CEF languages as possible, and include some large aligned parallel corpora for domains relevant to CEF DSIs[1], or for the general (administrative/regulatory) domain.

The data can be in **different formats**, including tmx, xml, xliff, html, doc, editable pdf etc. (Important note: Pdf should only be provided if the underlying original files in e.g. doc or another format are not available.)

[1] Online Dispute Resolution (ODR), Europeana, Open Data Portal, eJustice, Electronic Exchange of Social Security Information (EESSI)

1.4 How can I get help if I have any questions?

As indicated in 1.1 above, your **ELRC representative** is there to help if any questions arise regarding the **data collection** process. The list of the ELRC representative responsible for your country is available in Annex 1.

Moreover, for all **legal and technical questions** regarding the contribution of language resources, ELRC offers a central **helpdesk** to provide support to questions related to the provision and contribution of data. The helpdesk is available online through the ELRC website: http://www.lr-coordination.eu/helpdesk Simply visit us online!

Your questions can then be submitted

- through the Web forum (http://cef-at-helpdesk.elda.org/overview/) – any time
- by email (help@cef-at-helpdesk.org) – any time
- by phone (+33 970 440 522) – 9.30am-4.30pm CET from Monday to Friday
- via skype (CEF-AT-Helpdesk) – 9.30am-4.30pm CET from Monday to Friday

1.5 What is the best practice for formally engaging the public sector?

Ideally, communication should be driven by the National Anchor Point (in particular the **Public Service NAP**) in this particular country. Public Service NAPs are liaison contact persons to national, regional and local administrations. They have the mission to reach out to the target audience and spread the word about the importance of language resources and the ELRC effort among public authorities and ministries in their country.

As a typical means to encourage the public sector in your country, corresponding **letters or circulars from the relevant national authority** (typically Public Service NAP, but also other if appropriate) to suitable recipients can be sent. This letter/circular typically includes the invitation to collaboration, indicating details of the further process (e.g. details on how data could be provided, or even invitation to a joint meeting / web conference etc.). The communication strategy chosen should take into account the specialties of the administrative landscape and the particular situation in that country.

Support from the EC (e.g. official letter addressed to relevant national officials) can be provided where necessary. A corresponding request must be made to the ELRC representative in charge of this country (see Annex 1 for the list of the ELRC representative responsible for your country).

1.6 We do not have tmx/aligned corpora – can we send original translations?

If you do have any good translations in docx or pdf, in particular when translations are on domains relevant to CEF DSIs[2], you can provide them to us as well. As shown above, data can be contributed either via email (send email to data@lr-coordination.eu) or they can directly be uploaded into the ELRC repository through (http://www.lr-coordination.eu/resources). See section 1.2 as well as sections 2.1 - 2.3 for details.

1.7 If ELRC creates tmx files from our raw data, can we have them?

I would suggest yes, we should. This might be a good incentive for institutions to contribute relevant data.

[2] Online Dispute Resolution (ODR), Europeana, Open Data Portal, eJustice, Electronic Exchange of Social Security Information (EESSI)

1.8 Are you also interested in monolingual texts?

If you do have any good monolingual texts that are of terminological interest (i.e. in particular texts on domains relevant to CEF DSIs[3]) you can provide them to us as well. As shown above, data can be contributed either via email (send email to data@lr-coordination.eu) or they can directly be uploaded into the ELRC repository through (http://www.lr-coordination.eu/resources). See section 1.2 as well as sections 2.1 - 2.3 for details.

[3] Online Dispute Resolution (ODR), Europeana, Open Data Portal, eJustice, Electronic Exchange of Social Security Information (EESSI)

2 Uploading Data via the ELRC-SHARE Repository

2.1 The repository at a glance

The ELRC-SHARE repository is used for documenting (i.e. describing), storing, browsing and accessing LRs[4] that are considered useful for feeding the CEF.AT platform. All Language Resources (LRs) are documented with the ELRC-SHARE metadata (based on the META-SHARE schema).

The repository is accessible through the ELRC website (http://www.lr-coordination.eu/resources) and also directly via http://erlc-share.ilsp.gr (see below, Figure 1).

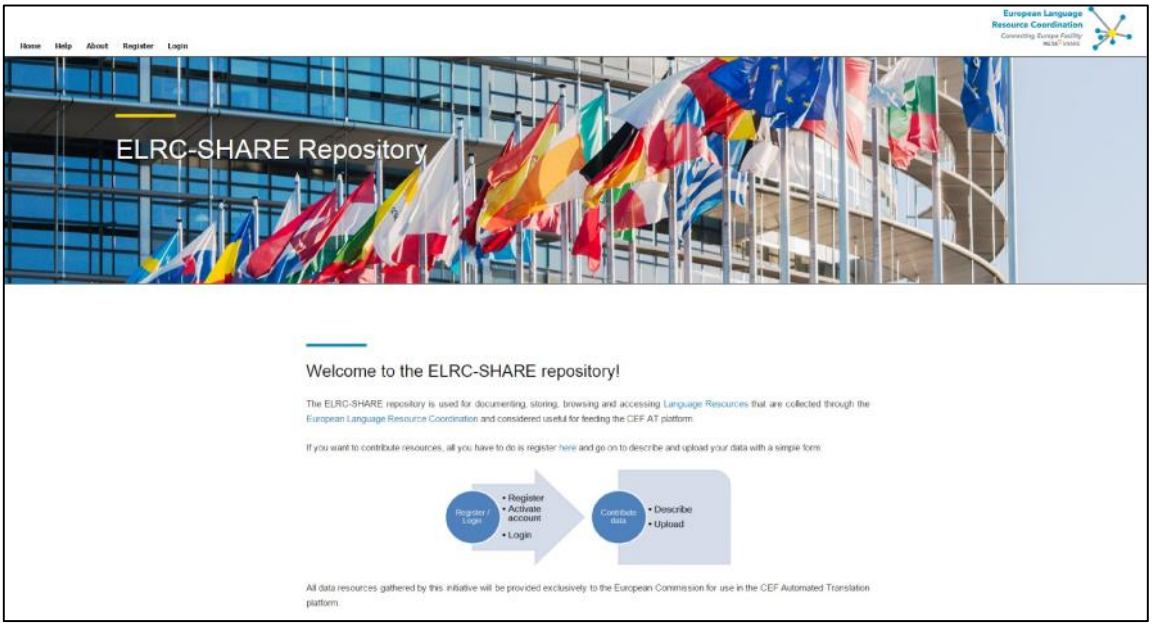
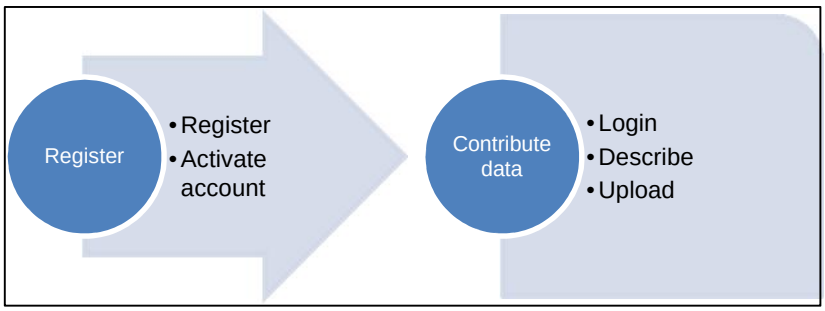


Figure 1 - LR inventory at the ELRC-SHARE repository

The process through which data contributors can upload data to the ELRC-SHARE repository is quite simple and involves the following two steps (see Figure 2 below):

- user registration or, if the user is already registered, login
- description and uploading of the data with a simple form.



[4] Data can only be browsed or assessed by the partners oft he ELRC consortium.

Figure 2 – General process for contributing data via ELRC-SHARE

2.2 Registration process with the ELRC-SHARE repository

You can contribute resources only as registered user. **Registration is essential for verifying the LR provider and avoiding abuse, and as such for keeping the repository and its contents secure, clean and monitored as to their use.**

To get registered, please go to the ELRC-SHARE repository either via the ELRC website (http://www.lr-coordination.eu/resources) or directly via elrc-share.ilsp.gr and simply click **Register** (see Figure 3 below).

Figure 3 – How to register with the ELRC-SHARE – Step 1 Registration Button

On the registration page which opens now, please fill in all the required information (see Figure 4 below).

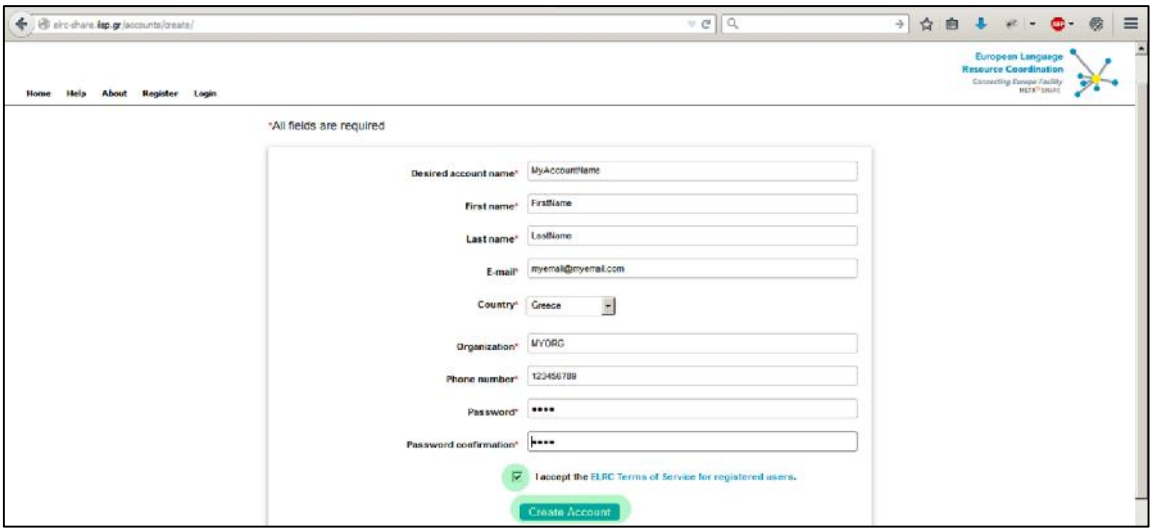


Figure 4 – How to register with the ELRC-SHARE – Step 2 User Details

Click on **ELRC Terms of Service for registered users** to read the terms of use for registered users. If you accept the terms of use, please check the checkbox next to **I accept the ELRC Terms of Service for registered users**. Then click the **Create Account** button.

A notification message appears that acknowledges receipt of the request and informs you that a confirmation email has been sent to the email address you have filled in at the registration form (see Figure 5 below).

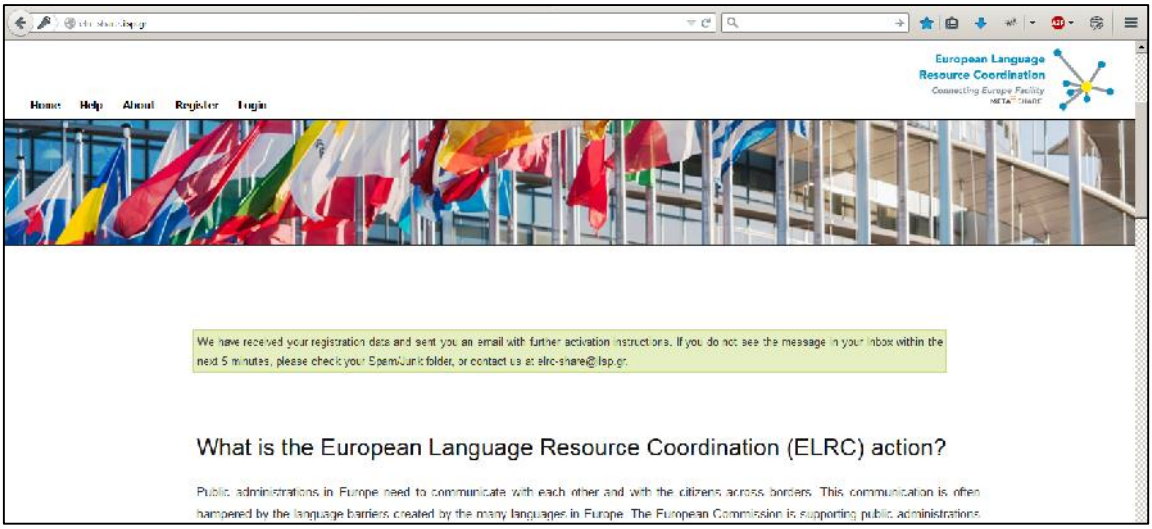


Figure 5 – How to register with the ELRC-SHARE – Step 3 Confirmation of your application

Check your email account and click the activation link that is included in the email message.

Your account will be activated and you will be directed to the **Data contribution form** (see Figure 6 below).

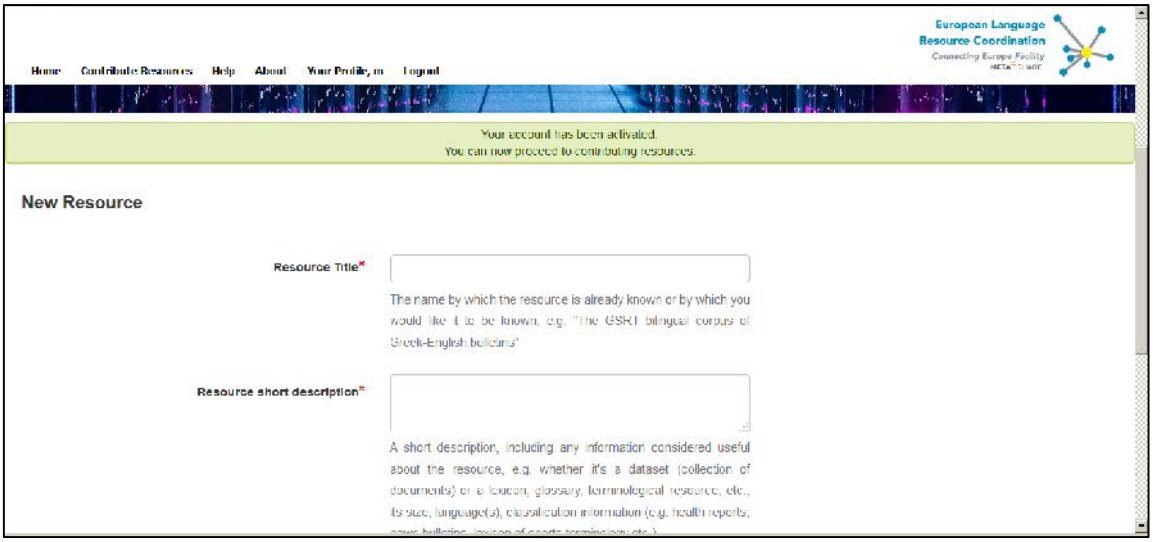


Figure 6 – How to register with the ELRC-SHARE – Step 4 Finalisation of registration

2.3 Process for contributing data via the ELRC-SHARE repository

As indicated above (see sections 2.1 and 2.2), all data contributors must be registered first (see section 2.2 for details of how to register). Data contributors can then upload and briefly document their data via a simple web form (Figure 7 below).

To access this form, simply click on **Contribute Resources** from the top menu.

Figure 7 – Contribute Resources Form

The form is kept very simple and includes fields to be filled in with information describing the resource (namely: resource title, resource short description, languages) and an **Upload** button for uploading the resource itself (see Figure 8 below). Fields with a red asterisk (*) are mandatory.

Figure 8 – Contributing Resources Web form at a glance

Fill in the appropriate information, click **Browse** to upload your resource. In the window that opens (see Figure 9 below), browse the folders of **your computer** for the respective zipped file (**.zip**) containing the data you want to contribute, select it and click **Submit**. Please note

that only zipped files up to 50 MB are allowed; if the file you try to upload exceeds this size, a warning message appears.

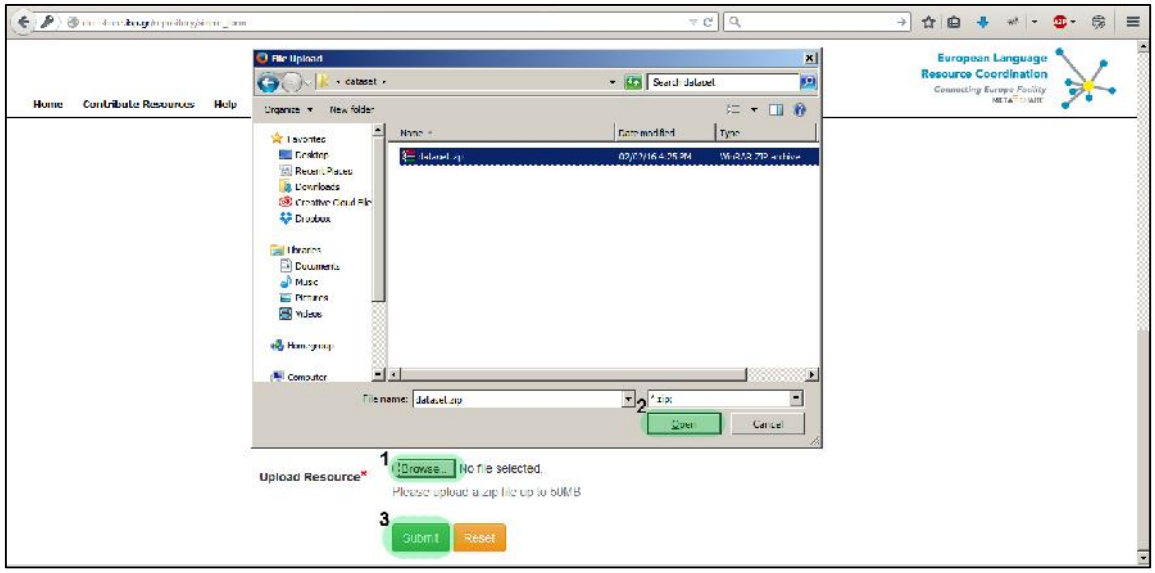


Figure 9 –Upload of resources using the web form

Upon submission of new data, the ELRC member(s) responsible for the contributor's country are notified by email. They can then log into the repository with their credentials and editor rights, where an initialized metadata record with all the information that the contributor has provided has been created.

The metadata record must be enriched with information that is required to conform to the ELRC-SHARE schema (e.g. domain, license, size and format of the resource) and, optionally, further information considered useful for the CEF.AT objectives (e.g. relevant DSI). Section 2.8 below provides details of the meta-data schema. To obtain this information, they will contact the data contributors to interview them and they can also inspect the uploaded data.

2.4 I work in the public sector – how can I donate LRs via the repository?

If you wish to contribute LRs via the ELRC-SHARE repository you must first register (see above, 2.2 for details); a valid email address is required for this process. Once you obtain the account credentials, you can log in, click on the menu item "contribute resources" and use a simple web form to describe and upload LRs (see section 2.3 for details).

For each LR, you must provide a very short description and a title in English and upload it in zipped format; the zip file can contain files in any document format, e.g. tmx, xml, xliff, html, doc, editable pdf etc. (Important note: Pdf should only be provided if the underlying original files in e.g. doc or another format are not available.)

Following the submission, you will soon be contacted by an ELRC member (see Annex 1 for ELRC representatives for data collection in each country) who will ask you for some minimal further information about the LR.

2.5 In which document format can I upload my language resource(s)?

You can upload any data only as **zip-file**. Language resources can be contributed in various formats, e.g. tmx, xml, xliff, html, doc, word-files, excel-files, editable pdf etc. (Important note: Pdf should only be provided if the underlying original files in e.g. doc or another format are not available.) Please also note that only zipped files up to 50 MB are allowed; if the file you try to upload exceeds this size, a warning message appears. If you are unable to provide a zipped file with less than 50MB, please contact the ELRC Helpdesk via email, phone or skype (see http://www.lr-coordination.eu/helpdesk for details).

2.6 How can I view and download LRs from the repository?

LRs contributed by data contributors must first undergo a process whereby the authorized editors enrich their descriptions and confirm/check the licensing conditions – for unclear cases, they can contact the legal helpdesk. During this process, the LR descriptions (metadata records) are considered "internal" and cannot be viewed by anyone else. When the documentation is finalized, the authorized editor will make the metadata record public through the inventory and the resource itself will become downloadable by users, in accordance with its licensing conditions.

2.7 What information should I provide with the language resource(s)?

Data contributors from the general public sector provide only a very short description and a title in English and, ideally, the language(s) of the LR. Authorized editors, typically ELRC members, are assigned the task of enriching these descriptions according to the ELRC metadata schema.

2.8 What is the ELRC-SHARE metadata schema?

The ELRC-SHARE schema includes metadata fields for the description of LRs, such as textual corpora, computational lexica, terminological resources, grammars etc. It includes common elements for all resources as well as distinct elements depending on the type of the resource. A subset of the elements is mandatory and this is clearly indicated on the editor.

The main metadata elements included in the schema are the following:

- Resource name, short name, identifier(s)
- Resource type:
  - corpus (e.g. monolingual corpus, bilingual corpora, translation memories etc.)
  - lexical/conceptual resource (e.g. dictionaries / lexica / ontologies / NE gazetteers)
  - language description (e.g. grammars)
- Contact person(s)
- Resource licence and conditions of use: each resource will be clearly identified either as Open Data, with the appropriate open licence, or as restricted/confidential language resources, specifying the licensing conditions and the IPR Holder(s), in

view of obtaining the right to use such restricted resources for setting up and adapting automated translation services for the CEF DSIs
- Resource size
- Resource format(s) (e.g. doc/tmx/xml etc.)
- Character encoding (e.g. UTF-8, ASCII etc.)
- Language(s)
- Resource creator(s)
- Domain(s), with values taken from the EuroVoc thesaurus (http://eurovoc.europa.eu/drupal/)
- Text classification, e.g. administrative texts, meeting proceedings etc.
- If applicable: Information on what pre-processing was done. If pre-processing was performed, the raw material must be provided, too.

2.9 Can public sector reps have access to the ELRC SHARE repository?

Data can be contributed by any person following the registration and upload process described above (see 2.2 for registration, 2.3 for data contribution). Data from the repository can currently only be browsed and accessed by ELRC. In particular, it is not possible by anybody else to download data. If there are any questions on the data contained in the ELRC SHARE repository, please do get in touch with the ELRC Helpdesk (see http://www.lr-coordination.eu/helpdesk for details).

3 Annex 1 – Your ELRC representatives

- **DFKI** (Andrea Lösch – andrea.loesch@dfki.de, Fraser Bowen – fraser.bowen@dfki.de):
  - Germany
  - Austria
  - Luxemburg
  - Netherlands
  - Hungary
  - Czech Republic

- **ELDA** (Khalid Choukri – choukri@elda.org, Hélène Mazo – mazo@elda.org):
  - U.K.
  - Ireland
  - Spain
  - Portugal
  - Belgium
  - Italy
  - Malta
  - France

- **Tilde** (Aivars Berzins - aivars.berzins@tilde.lv, Roberts Rozis - roberts.rozis@tilde.com):
  - Latvia
  - Estonia
  - Lithuania
  - Finland
  - Sweden
  - Denmark
  - Iceland
  - Norway

- **ILSP** (Kanella Pouli – kanella@ilsp.gr, Penny Labropoulou – penny@ilsp.gr):
  - Greece
  - Cyprus
  - Slovakia
  - Slovenia
  - Bulgaria
  - Poland
  - Romania
  - Croatia

4 Annex 2 – Potential Data Holders

4.1 Potential Data Holders Greece

Organisation
Bank of Greece
Civil Aviation Authority
Freelance translator
General Secretariat for Industry
Greek Free/Open Source Software Society
Hellenic Army
Hellenic Ministry of National Defence/General Directorate of Defence Policy & International Relations
Hellenic Police
Hellenic Single Public Procurement Authority
Hellenic Statistical Authority
Hellenic Telecommunications Organization
Institute of Eastern Mediterranean
Institute of National Relations
Managing Authority "Reform of the Public Sector"
Ministry of Economy
Ministry of Foreign Affairs
Ministry of Health
Ministry of Interior and Administrative Reform
Ministry of Justice
Ministry of Labour
Municipality of Acharnai
National Documentation Centre
National Museum of Contemporary Art
National School of Public Administration
NGO of the Archbishopric of Athens
Special Secretariat for Water
Technological Educational Institute of Athens
The Media Institute, UCL
The National Centre for Public Administration and Local Government
TMServe
Union of the Graduates of the School of Public Administration (ENAP)

4.2 Potential Data Holders Germany

Auswärtiges Amt	3
Bundesinstitut für Berufsbildung	1
Bundesministerium der Verteidigung	2
Bundesministerium des Innern	2
Bundesministerium für Arbeit und Soziales	1
Bundesministerium für Finanzen	1
Bundesministerium für Gesundheit	1
Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit	1
Bundesministerium für Verkehr und digitale Infrastruktur	1
Bundesministerium für Wirtschaft und Energie	2
Bundesnetzagentur	1
Bundessprachenamt	3
Bundesverband der Dolmetscher und Übersetzer	1
Deutsche Bahn	2
Deutscher Bundestag	1
DFKI GmbH	6
DGT Local Field Office Berlin	1
DIN e.V.	1
Europäische Kommission	1
European Language Resources Association	2
Fraunhofer-Gesellschaft	1
GlobalSprachTeam	2
Institut für Deutsche Sprache	1
Institute for Language and Speech Processing, "Athena" RC	1
iRights	1
Karlsruhe Institute of Technology	2
Ministerium für Wirtschaft und Energie Brandenburg	1
Rat für Sozial- und Wirtschaftsdaten	1
Senatskanzlei Berlin	1
Staatskanzlei des Saarlandes	1
Volkswagen AG	1
ZF Friedrichshafen AG	2

4.3 Potential Data Holders Latvia

Academic Information Centre	2
Culture Information System Centre (KISC)	1
Latvian Academy of Sciences Terminology Commission	1
Latvian Association of Local and Regional Governments	1
Latvian Geospatial Information Agency	4
Latvian Parliament Saeima	3
Ministry of Education and Science	1
Ministry of Environmental Protection and Regional Development of Latvia, Terminology Commission	1
Ministry of Health	1
Sportacentrs (sports news website)	1
State Education Development Agency	1
State Employment Agency	1
State Regional Development Agency	1
The Latvian Institute	1
The Latvian Language Agency	2
The Latvian National Standardisation Body Latvian Standard (LVS)	1
The Latvian State Language Center	5
The Ministry of Environmental Protection and Regional Development	1
The National Archives of Latvia	2
The Office of Citizenship and Migration Affairs	3
The State Chancellery	1
Translate 24/7	1
University of Latvia Agency "Latvian Language Institute of the University of Latvia	2