Preparing and sharing data with the ELRC-SHARE repository and what happens next

Maria Giagkou
Institute for Language and Speech Processing / Athena R.C.
ELRC
The notion of data

- Weather / climate
- Geospatial
- Educational
- Public spending
- Demographic
- Public health and insurance
The notion of data

Data: the oil of the 21st century
The notion of data
The notion of language data

Data

- any piece of electronically stored content

(Textual) Language Data

- any piece of electronically stored text
Očuvana priroda doprinosi osiguravanju svih funkcionalnosti nužnih za život i ekonomski razvoj. U Republici Hrvatskoj, kao i u svijetu, priroda je pod stalnim pritiskom ljudskih djelatnosti. Iako se ulažu značajniji napori za očuvanje prirode, pojedine njezine sastavnice su i dalje ugrožene.

Conserved nature contributes to ensuring all functionalities necessary for livelihoods and economic development. In the Republic of Croatia, as well as in the world, nature is facing permanent pressures from human activities. Even though significant efforts are being invested in nature conservation, certain nature components are still being threatened.
Data used by eTranslation

DGT translation memory

Council of the European Union

European Commission

CVRIA

European Parliament

European Committee of the Regions

European Investment Bank

European Central Bank
Such data are already available  
BUT  
they are not enough...
What data are useful for eTranslation as per type

• Any **electronically stored text** in an EU language plus NO and IS
• **Texts and their translations** (i.e. parallel bilingual or multilingual)

---

**Croatian text**


---

**Translation in English**

In this Statistical Report, the Croatian Bureau of Statistics puts at users' disposal the final data of the Census of Population, Households and Dwellings 2011. This Statistical Report presents data on age and sex by counties, towns, municipalities, settlements and the districts of City of Zagreb. Data are presented according to the situation as on 31 March 2011 at midnight (census moment) by the then relevant territorial constitution. The Census of Population, Households and Dwellings 2011 was carried out in the Republic of Croatia from 1 to 28 April 2011 according to the situation as on 31 March 2011 at midnight, which is deemed to be the census moment.
• List of terms and their translations, i.e. a **terminology**, e.g.

<table>
<thead>
<tr>
<th>Croatian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerodinamika</td>
<td>aerodynamics</td>
</tr>
<tr>
<td>zračna struja</td>
<td>air flow</td>
</tr>
<tr>
<td>zračna struja</td>
<td>air stream</td>
</tr>
<tr>
<td>granični sloj</td>
<td>boundary layer</td>
</tr>
<tr>
<td>čeoni otpor</td>
<td>drag</td>
</tr>
<tr>
<td>koeficijent otpora</td>
<td>drag coefficient</td>
</tr>
<tr>
<td>dinamički tlak</td>
<td>dynamic pressure</td>
</tr>
<tr>
<td>slobodna struja zraka</td>
<td>free air stream</td>
</tr>
<tr>
<td>aerodinamički uzgon</td>
<td>aerodynamic lift</td>
</tr>
<tr>
<td>koeficijent uzgona</td>
<td>lift coefficient</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
• In principle, any text in machine readable format
• But, some formats are more “MT-ready” than others, i.e. they require less manual or automatic processing
• More processing introduces more errors in the final output, making it less useful for eTranslation
Minority Rights in Croatia

Dario Kuntić

The aim of this paper is to present and to explain the level of protection of minority rights in the Republic of Croatia. As Croatia inherited from the former Socialist Federal Republic of Yugoslavia the system of the protection of rights of the minorities immediately after gaining independence, it provided a model of the realization of minority rights. That model guaranteed minorities the right for education in their own language, the right to the official use of their language and alphabet, the right to publishing in their mother tongue, access to the media, protection of cultural treasures and their entire cultural heritage, various venues of preservation of ethnic, language and religious identity and the representation of minority interests.

Key words: Republic of Croatia, minorities, minority rights, democratic principles

1. Introduction

By its Constitution, the Republic of Croatia bounds itself to protection of the minorities from any discrimination. They have the right to identity and participation in public affairs. Members of ethnic and national communities or minorities whose share in the total population is higher than 3% have the right for participation, proportional with their share in the entire population, in the Croatian Parliament, the Government of the Republic of Croatia and in the bodies of supreme judicial authorities.

Members of national minorities exercise their political rights, like all citizens or nationals in the Republic of Croatia through regular institutions on the basis of the provisions of the Constitution, Constitutional Law, Law on Election of Representatives to the Croatian Parliament, Law on Election of Members of Representation Bodies of Units of Self-government and Administration and Law on Political Parties.

The Republic of Croatia signed a number of international, bilateral and multilateral, contracts and agreements related to the protection of human and minority rights such as the European Convention on Human Rights, the European Charter of Local Self-government, Charter on Regional and Minority Languages, the Framework Convention for the Protection of National Minorities, the Treaty on the Protection of the Italian Minority in the Republic of Croatia and the Croatian Minority in the Republic of Italy, and many others.

At the end of this introduction it would be important to mention that according to the 1991 census in Croatia there were 12,012 Albanians, 214 Austrians, 43,469 Muslims, 458 Bulgarians, 9,724 Montenegrins, 13,049 Croats, 22,355 Hungarians, 6,250 Macedonians, 2,655 Germans, 679 Poles, 8,695 Roma, 810 Romanians, 706 Russians, 3,506 Slovaks, 581,063 Serbs, 21,303 Italians, 330 Turks, 2,694 Ukrainians, 22 Vlachs, 600 Jews and 3,012 other ethnic and national minorities.

* Dario Kuntić is a political scientist working in the Ministry of European Integration of the Republic of Croatia and holds an M.A. in Democratization and Human Rights from the University of Bologna, Italy.

Original paper
UDC 323.15(497.5)
Received in May 2003
The following formats are particularly useful (in descending order):

- For bilingual/multilingual parallel texts
  1. Translation memories (.tmx)
  2. XML translation files (.xliff)
  3. Plain text (.txt, .csv)
  4. Spreadsheets (e.g. xlsx)

- For terminologies
  1. TermBase eXchange (.tbx)
  2. Plain text (.txt, .csv)
  3. Spreadsheets (e.g. xlsx)

- For monolingual texts
  1. Plain text (.txt, .csv)
File formats of parallel texts and their manipulation
Don’ts

This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.


A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English.

Ovo je drugi stavak na hrvatskom jeziku. Ovo je drugi stavak na hrvatskom jeziku. Ovo je drugi stavak na hrvatskom jeziku. Ovo je drugi stavak na hrvatskom jeziku.
## Don’ts

<table>
<thead>
<tr>
<th>English</th>
<th>Hrvatski</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.</td>
<td>Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku. Ovo je odlomak lijevo na hrvatskom jeziku.</td>
</tr>
</tbody>
</table>
Preparing your data | 3

Use **identical filenames** for each document pair (source – translation)
Include language identifiers in the filename
Criteria for grouping your data into meaningful datasets

• A dataset is a collection of data **grouped according to certain criteria**

• For the purpose of enhancing and adapting CEF eTranslation, two criteria are critical:
  
  – **Language(s)**: each collection is defined by the language or language pairs of its data, e.g.
    
    • *Collection of texts in English – Croatian*
    • *Documents in English – Croatian - French*
  
  – **Domain**: each collection ideally belongs to a single domain, e.g.
    
    • *Collection of texts in English – Croatian in the culture domain*
    • *Social security documents in English – Croatian - French*
• Administrative/regulatory domain and
• Topics relevant to the CEF DSIs

<table>
<thead>
<tr>
<th>CEF DSI</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Dispute Resolution</td>
<td>Consumers’ rights, complaints</td>
</tr>
<tr>
<td>Electronic Exchange of Social Security Information</td>
<td>Social security, insurance</td>
</tr>
<tr>
<td>eProcurement</td>
<td>Public procurement, contractual agreements</td>
</tr>
<tr>
<td>European e-Justice Portal</td>
<td>Justice, Law</td>
</tr>
<tr>
<td>eHealth</td>
<td>Health, Medicine</td>
</tr>
<tr>
<td>Business Registers Interconnection System</td>
<td>Business, market</td>
</tr>
<tr>
<td>Safer Internet</td>
<td></td>
</tr>
<tr>
<td>Cybersecurity</td>
<td></td>
</tr>
<tr>
<td>Public Open Data</td>
<td></td>
</tr>
<tr>
<td>Europeana</td>
<td>Culture</td>
</tr>
</tbody>
</table>
How to contribute your data to CEF eTranslation
A step-by-step guide
The ELRC-SHARE Repository

- Access to, sharing and contribution of language resources
- Access to tools and services catalogue

https://elrc-share.eu
How to Contribute Data

Register
- Register
- Activate account

Contribute Data
- Login
- Describe
- Upload
How to Register (1/2)

Welcome to the ELRC-SHARE repository!
How to Register (2/2)

• Fill in the required info
• Read the Terms of Service and click Accept, if you agree
• Click the Create Account button
• Activate your account according to the guidelines emailed to you
How to Contribute Data (1/6)

Data Contribution

New Resource

Resource Title*

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"
How to Contribute Data (2/6)

- Fill in the details of the dataset

![Form for filling in resource details](image)
How to Contribute Data (3/6)

• Three modes for contributing your data

**Contribution Mode**
- Upload ZIP archive
- Provide URL of resources
- eDelivery (Generate XML file to attach to your eDelivery contribution)

Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

**Upload Resource**

Choose File  No file chosen

Please upload a *.zip file up to 100MB.

In case the *.zip file file you wish to upload is larger than 100MB, please contact elrc-share@ilsp.gr
1. Click on Choose file
2. Locate your resource in your hard disk
3. Click on Submit
• Alternatively indicate a url (directory listing)

Language(s)
- Bulgarian
- Czech
- Croatian
- Danish
- Dutch; Flemish
- English
- Estonian
- Finnish
- French
- German
- Hungarian

The language(s) of the resource; for resources with multiple languages, hold down CTRL key to select multiple values

Contribution Mode
- Upload ZIP archive
- Provide URL of resources

Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

Resource URL

Please provide a URL containing the files you wish to contribute

Submit  Reset
Contribute Your Data Through eDelivery

If you wish to share your data through eDelivery, you can use the ELRC-SHARE CEF eDelivery Access Point. For more information click here. In such a case, please fill in the form on the left and choose eDelivery in the Contribution mode.
Guidelines for contributors

Documentation on the ELRC-SHARE editor

The following guidelines provide detailed information on how to use the editing facility for documenting and uploading LRs:

- Walkthrough for contributors
- Walkthrough for editors

ELRC-SHARE schema

- ELRC-SHARE schema XSD (based on the META-SHARE Schema)
- Documentation about the schema
Preparing and sharing data with the ELRC-SHARE repository

and what happens next
What happens to your data?

Data contributor

Upload to ELRC-SHARE

ELRC processes your data

Processed data

CEF Digital
Connecting Europe
Language processing services

Data extraction
If your data is trapped in archives and databases, we can help extract it

Anonymisation
Does your data contain private info? We can help to anonymise

Cleaning
If your data is messy (i.e., lots of noise), we will clean it up

Re-formatting
Need to re-format DOCX to XML, or PDF to WORD? Let us do it for you!
Language processing services

Data conversion
If your data isn’t converted to the proper formats, we can help convert it.

Tag removal
Does your data contain unneeded tags? We can assist in removing them!

Alignment
Translations aren’t aligned? We’ll do it for you with our tools!

Metadata
Metadata are crucial! We can organise and validate metadata for your team.
After processing:

File01_hr.txt
File01_en.doc
File02_hr.pdf
File02_en.txt
File03_hr.doc
File03_en.doc

What has happened to your data?

`<tu tuid="818">
  <tuv xml:lang="hr">
    <seg>U Republici Hrvatskoj, kao i u svijetu, priroda je pod stalnim pritiskom ljudskih djelatnosti.</seg>
  </tuv>
  <tuv xml:lang="en">
    <seg>In the Republic of Croatia, as well as in the world, nature is facing permanent pressures from human activities.</seg>
  </tuv>
</tu>`

`<tu tuid="819">
  <tuv xml:lang="hr">
    <seg>Iako se ulažu značajniji napori za očuvanje prirode, pojedine njezine sastavnice su i dalje ugrožene.</seg>
  </tuv>
  <tuv xml:lang="en">
    <seg>Even though significant efforts are being invested in nature conservation, certain nature components are still being threatened.</seg>
  </tuv>
</tu>`
How your dataset is described
All these services can also be offered **on-site** to all data contributors **free of charge**
Our team of experts will travel directly to assist you at your own offices
On-site assistance

We will fix your data issues and return the processed data directly to you. We can also help to improve your data management processes. Just ask!
Data contributor

Upload to ELRC-SHARE

ELRC processes your data

Processed data

Request on-site assistance
How to request services and help
Submit a request for on-site assistance by filling out the form below. See a list of services here.

First name *

Last name *

Institution *

Country *

Email *

Types of assistance required *
- Legal assistance
- Data processing
- Anonymisation
- Other

Description of assistance required

Submit
ELRC Helpdesk

www.lr-coordination.eu/helpdesk

Helpdesk for Language Resources

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

<table>
<thead>
<tr>
<th>Channel</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone*</td>
<td>+33 970 440 522</td>
</tr>
<tr>
<td>Secretariat Support</td>
<td>+49 681 857 7552 85</td>
</tr>
<tr>
<td>Skype</td>
<td>ELRC Helpdesk</td>
</tr>
<tr>
<td>E-mail</td>
<td><a href="mailto:help@lr-coordination.eu">help@lr-coordination.eu</a></td>
</tr>
</tbody>
</table>
Hvala vam!
Icons used in this presentation

- By Michael Mellon, GB, CC-BY 3.0 US
- By Joana Pereira, BR, CC-BY 3.0 US
- By Becca O'Shea, NZ, CC-BY 3.0 US
- By Creative Stall, Basic licence www.iconfinder.com
- By Creative Stall, PK, CC-BY 3.0 US
- By Arthur Shlain, IL, CC-BY 3.0 US
- By Shmidt Sergey, US, CC-BY 3.0 US
- By Gregor Cresnar, CC-BY 3.0 US
- By anbileru adaleru, CC-BY 3.0 US
- By Vectors Market, CC-BY 3.0 US