

**EUROPEAN
LANGUAGE
GRID**



Sharing datasets & models through ELG

Penny Labropoulou (ILSP/Athena RC)

15/16/17-11-2021 META-FORUM 2021 – Using the European Language Grid (virtual conference)

<http://www.european-language-grid.eu>

Registering your resource

The screenshot displays the European Language Grid (ELG) user interface. At the top left is the ELG logo with the text 'EUROPEAN LANGUAGE GRID' and 'RELEASE 2'. The top navigation bar includes links for 'Technologies', 'Resources', 'Community', 'Events', 'Documentation', and 'About ELG'. On the right, there are links for 'My grid' and 'Test Provider'. Below this is a teal header bar with 'My grid', 'My items', 'Feedback', and 'Go to catalogue'.

The main content area is divided into several sections:

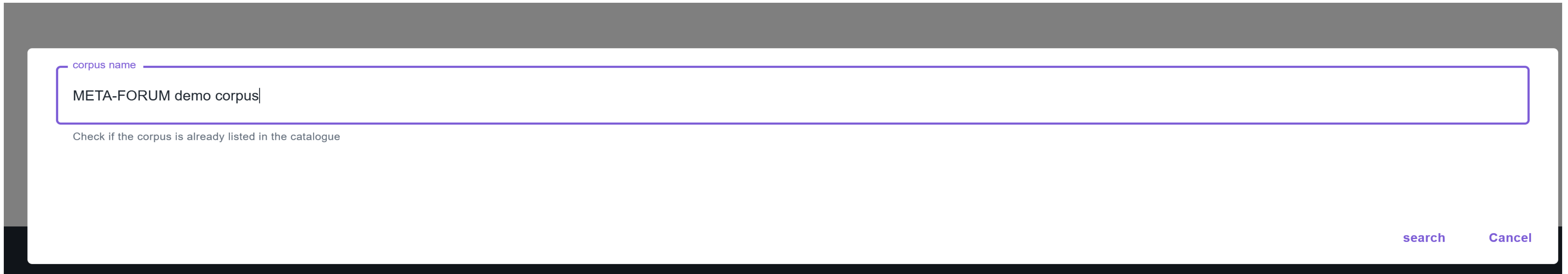
- Test Provider:** penny2spy@gmail.com, with a 'View profile' button.
- Welcome to your grid !:** A section titled 'Here you can:' with a list of actions: view and update your profile, view your items and your tasks, access creation forms, and upload items.
- Total items:** A card showing 'Number of items you have created.' with the value '549' and a '+ Create items' button.
- Upload items:** A card with the text 'Upload single or multiple items in XML format.' and a '+ Upload items' button.
- Validate your XML:** A card with the text 'Validate your XML before uploading.' and a '+ Validate XML' button.

Red circles are drawn around the '549' in the 'Total items' card, the '+ Upload items' button, and the '+ Create items' button.

- Same as for services: you can
 - upload single or batch XML files or
 - use the interactive editor

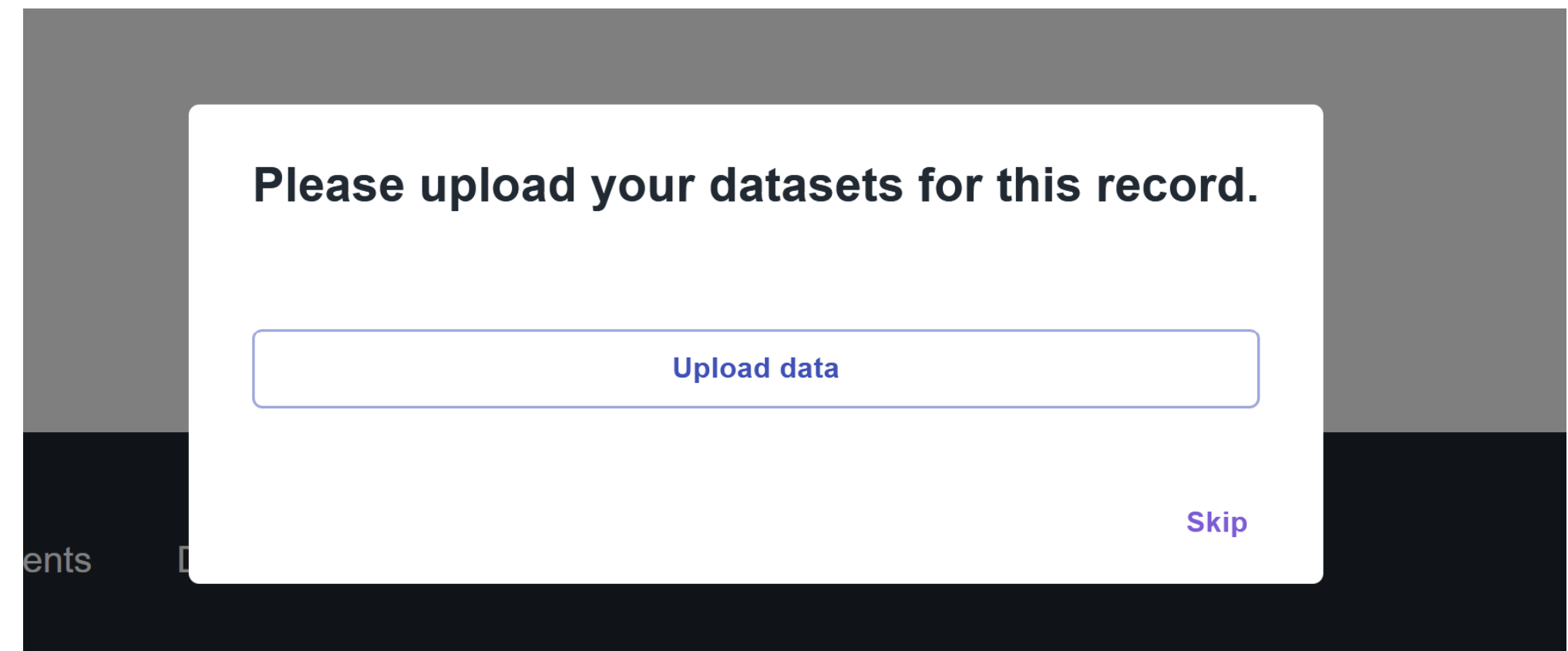
Step1: Create the metadata record

- As for services, start by entering a name for your resource, and do an initial check that it hasn't already been registered (by you or someone else!)



A screenshot of a web form for creating a metadata record. It features a text input field labeled 'corpus name' containing the text 'META-FORUM demo corpus'. Below the input field, a small text label reads 'Check if the corpus is already listed in the catalogue'. At the bottom right of the form, there are two buttons: 'search' and 'Cancel'.

- You will be prompted to upload your files; don't worry if you don't have it ready yet; you can start creating the metadata and upload it later



A screenshot of a modal dialog box with the title 'Please upload your datasets for this record.' Inside the dialog, there is a button labeled 'Upload data'. At the bottom right of the dialog, there is a link labeled 'Skip'.

Step 2: Basic information common to all resources

- The first tab (“Language resource/technology”) is the same as for services, i.e. it includes all the general / administrative metadata
 - Identity: **name**, **description**, **version**, logo, links to provider and funding project
 - Categories
 -

	CORPUS	PART	DISTRIBUTION	DATA	<input type="checkbox"/> Work in progress	Save draft	Save
IDENTITY CATEGORIES CONTACT DOCUMENTATION RELATED LRTS	LRT name * <input type="text" value="META-FORUM demo corpus"/> <small>The official name or title of the language resource/technology</small>		language <input type="text" value="English"/> <small>select language</small>				
	LRT identifier <input type="text"/> <small>A string used to uniquely identify the language resource/technology</small>				<button>Fill in</button>		
	LRT short name <input type="text"/> <small>An abbreviation, acronym, etc. used for the language resource/technology</small>		language <input type="text" value="English"/> <small>select language</small>				
	Description <div> Paragraph ▼ B I U ↶ ↷ ≡ ≡ ≡ ≡ <> ▼ </div> <input style="height: 150px;" type="text"/>		language <input type="text" value="English"/> <small>select language</small>				
	Version <input type="text" value="1.0.0 (automatically assigned)"/> <small>Recommended format: major_version.minor_version.patch (see semantic versioning guidelines at http://semver.org)</small>						
Version date <input type="text"/> <small>The date of the LRT version (latest update of the particular version if possible)</small>							
LRT provider							

Step 3: Information specific to corpora

- Corpus tab is for corpora irrespective of the media parts (text, audio, etc.)
 - Technical:
 - corpus subclass
 - GDPR-related information

The screenshot displays the 'CORPUS' tab in a web application. The top navigation bar includes tabs for 'LANGUAGE RESOURCE/TECHNOLOGY', 'CORPUS' (active), 'PART', 'DISTRIBUTION', and 'DATA'. On the right of the top bar, there is a 'Work in progress' checkbox and 'Save draft' and 'Save' buttons. A left sidebar contains a 'TECHNICAL' icon. The main content area has a 'Corpus subclass *' dropdown menu. Below it, a note reads: 'Select the type of the corpus: 'raw' for non-processed corpora, 'annotated' for corpora that include both the raw corpus and the processed output, 'annotations' for corpora that consist only of the processed output'. The form then asks 'Personal data included *' with 'Yes' and 'No' radio buttons, followed by a 'Specify if personal data are included' label. Next is 'Sensitive data included *' with 'Yes' and 'No' radio buttons, followed by a 'Specify if sensitive data are included' label. Then, 'Anonymized' with 'Yes' and 'No' radio buttons, followed by a 'Specify if the corpus has been anonymized' label. At the bottom right, there are 'Save draft' and 'Save' buttons.

Step 4: Information specific to "media parts" of a corpus

- First select the appropriate media type to reveal the information you can add
- Note: for multimedia corpora (i.e. corpora composed of parts with different media types, first describe one media part and you will be able to add more at the end

The screenshot shows a web interface for managing a corpus. At the top, there is a teal header bar with links for 'My grid', 'My items', 'Feedback', and 'Go to catalogue'. Below this is a navigation bar with tabs: 'LANGUAGE RESOURCE/TECHNOLOGY', 'CORPUS', 'PART' (which is highlighted in purple), 'DISTRIBUTION', and 'DATA'. To the right of these tabs are a 'Work in progress' checkbox and 'Save draft' and 'Save' buttons. The main content area is titled 'MEDIA PART' with a sub-header 'Corpus part' and a description: 'Describe each corpus part with a different media type (text, audio, video, etc.) separately'. A dropdown menu labeled 'Corpus media type' is open, showing options: 'text part', 'numerical text part', 'audio part', 'video part', and 'image part'. Below the dropdown, there are two buttons: 'Remove text part' and 'Add another Corpus part'. At the bottom right, there are 'Save draft' and 'Save' buttons.

Step 4: Information specific to "media parts" of a corpus

- Depending on the media type, you will be asked to fill in different information
- Common information for all: **language**
- Other types: classification information (e.g. text type)

The screenshot shows a web interface for managing language resources. At the top, there are tabs: 'LANGUAGE RESOURCE/TECHNOLOGY', 'CORPUS', 'PART' (which is active), 'DISTRIBUTION', and 'DATA'. To the right of these tabs are buttons for 'Work in progress', 'Save draft', and 'Save'. Below the tabs, on the left, is a sidebar with a 'MEDIA PART' icon. The main content area is titled 'Corpus part' and includes a subtitle: 'Describe each corpus part with a different media type (text, audio, video, etc.) separately'. The form is for a 'text part' and includes a 'Remove text part' button. The form fields are: 'Linguality type' (set to 'monolingual'), 'Language' (set to 'English'), 'Script', 'Region', 'Variant identifier', and 'Language variety'. There is also a 'language' dropdown set to 'English' with a '+' button next to it. An 'Add' button is at the bottom left of the form.

LANGUAGE RESOURCE/TECHNOLOGY CORPUS **PART** DISTRIBUTION DATA ☐ Work in progress [Save draft](#) [Save](#)

MEDIA PART

Corpus part
Describe each corpus part with a different media type (text, audio, video, etc.) separately

text part [Remove text part](#)

Linguality type *
monolingual
Select a value to indicate whether this corpus (part) includes one, two or more languages

Language
The language of the contents of the corpus (part) [Remove](#)

Language *
English
Start typing to select the language

Script
Start typing to select the script

Region
Start typing to select the region

Variant identifier
Start typing to select the variant (selection of values depends on the language)

Language variety
The name of the language variety (e.g. dialect), if applicable

language
English
select language [+](#)

[Add](#)

Step 4: How to access the resource

- Distribution tab
 - **Distribution form**: e.g. downloadable, accessible through interface
 - Media-dependent features: **size**, **format**, ...
 - **Licensing**: licence and cost

Licence

Start typing to select the licence of the corpus distribution, or add a new value

Licence name *

Cost

The cost for using the corpus

Fill in

Access rights

The rights for accessing the distributable form(s) of the corpus (preferably in accordance to a formalised vocabulary)

Fill in

LANGUAGE RESOURCE/TECHNOLOGY CORPUS PART **DISTRIBUTION** DATA ☐ Work in progress [Save draft](#) [Save](#)

Dataset distribution 1 [Remove Dataset distribution](#)

Describe separately each distributable form of the corpus (e.g., downloadable form in CSV, XML formats, form accessible via an iif)

Dataset distribution form *

Select the form or delivery channel through which the corpus is distributed

Private
☐ Yes
☐ No
Select "yes" if you want the access/download location to be hidden at the public metadata record

Samples location [Browse](#) [+](#)

A URL with samples of the corpus distribution; you can also upload a file

Text features [Remove](#)

The set of features (format, size) that describes this text distribution

Size
The size of the corpus distribution expressed as an integer and a unit value

Amount **Size unit**

Amount * Size unit * [×](#) [+](#)

The size of the corpus distribution Select the unit for the size

Data format

Data format *

Select the format of the language description distribution (multiple values possible)

Character encoding

Select the character encoding of the corpus distribution

[Add](#)

Step 4: Upload the dataset

- Data tab: Click to upload data

LANGUAGE RESOURCE/TECHNOLOGYCORPUSPARTDISTRIBUTIONDATA

☐ Work in progress

Save draftSave

DATA

Upload data

- Upload the file (zip files only allowed)

Please select a .zip file in order to upload a dataset for this record

Drag & Drop your files or [Browse](#)

cancelupload dataset

- When completed, you will see the following screen

LANGUAGE RESOURCE/TECHNOLOGYCORPUSPARTDISTRIBUTIONDATA

☐ Work in progress

Save draftSave

DATA

Upload data

Name	Upload date	Assigned to distribution	Actions
test1.zip	17 November 2021	No	

*In order to delete a dataset you should first unlink it from the corresponding distribution and save your record.

Save draftSave

Step 5: Link the dataset to a distribution

- Go back to the distribution tab
- If the distribution form is set to "downloadable", you can associate the dataset with the distribution

Private

- ☐ Yes
- ☐ No

Select "yes" if you want the access/download location to be hidden at the public metadata record

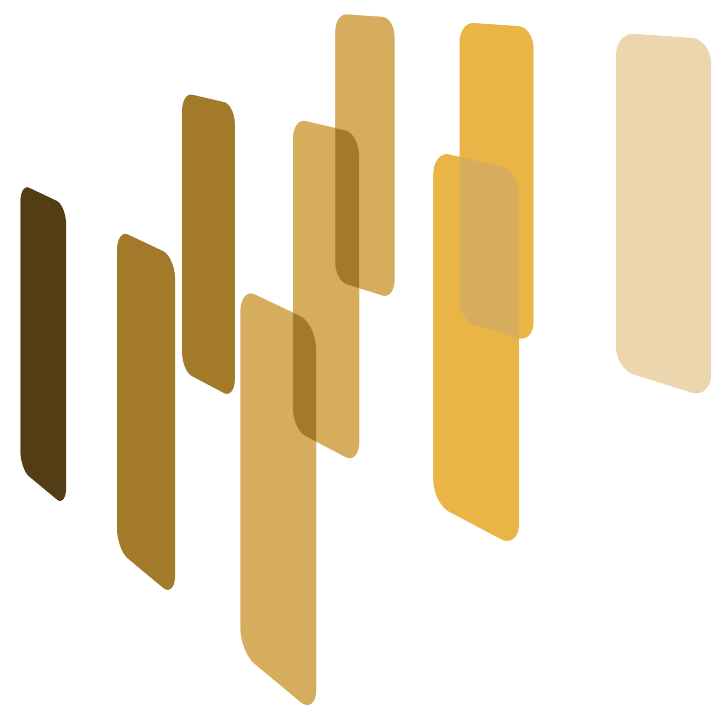
Associate a dataset with this distribution

test1.zip

- Yes, this is a complex procedure, but we are faced with different requirements for different applications

And finally...

- As for services, when you are done, submit the record for publication and we will contact you for any issues
- The process for sharing models and lexical resources is the same but with information specific to each resource type
- More information can be found at:
 - <https://european-language-grid.readthedocs.io/>
 - <https://www.youtube.com/channel/UCarEHmsWT2JslcvvWkbhL4A> (tutorials to be added)
- At any time, if something is unclear or you have any doubts, don't hesitate to contact us:
<https://live.european-language-grid.eu/catalogue/feedback>



European Language Grid

Thank you!

Time for discussion and questions



The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825627 (ELG).