European Language Resource Coordination
Connecting Europe Facility

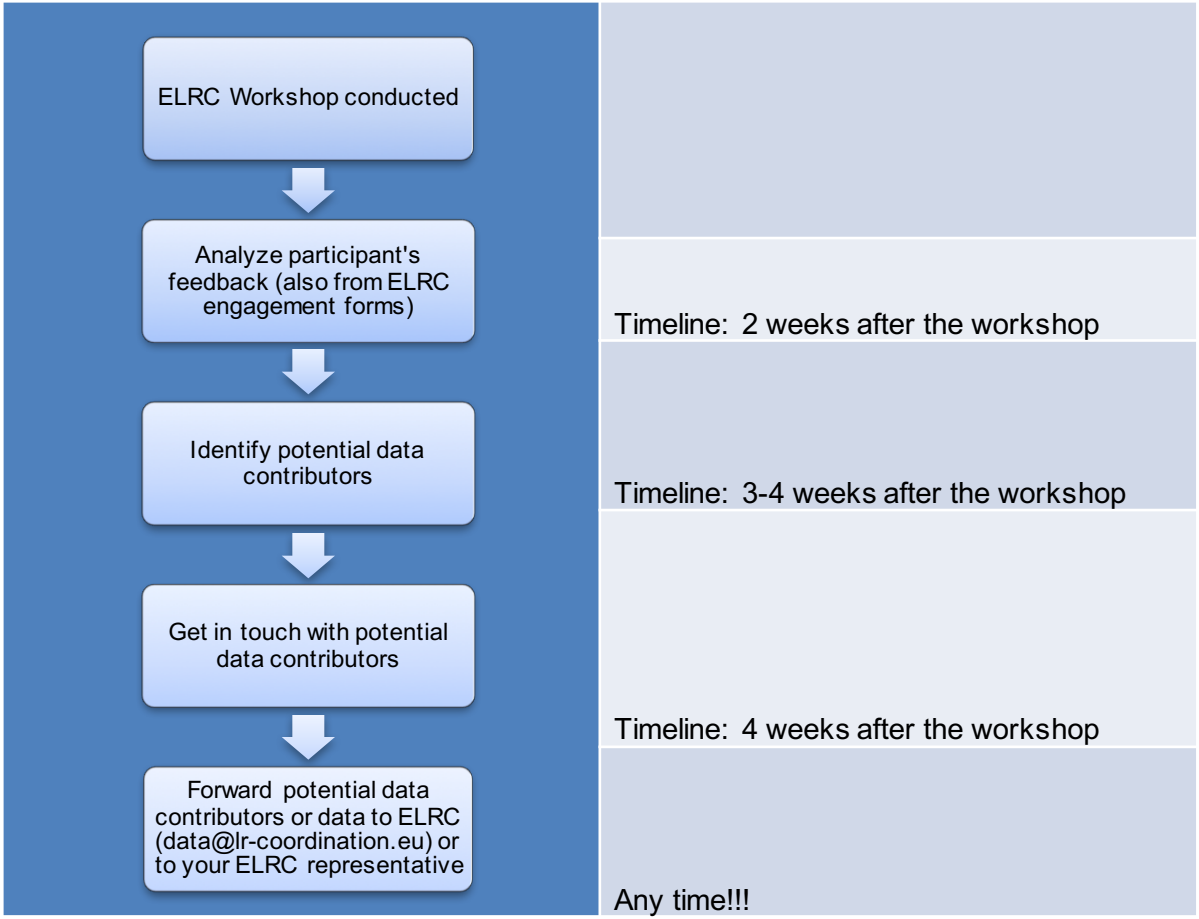
ELRC – Introduction to Data Collection

Josef van Genabith

Andrea Lösch

European Language Resource Coordination
Connecting Europe Facility

General process for data collection



European Language Resource Coordination
Connecting Europe Facility

Key steps in each country

- Monthly Q&A Online Sessions since May 2016 for each region
- To do's (until mid-July)
  - Finalisation of official letters of support (where needed)
  - Identification of potential data holders
    - Starting point: ELRC Engagement Forms
    - Continuous update, s. ELRC Country Overview
  - Decision on data collection strategy / contacting data holders
  - Signature of subcontract (where wanted)
  - Individual follow-up with data holders
- Important note: At least 6 potential data holders are needed per country, also as invitee for the ELRC Conference in October!!

European Language Resource Coordination
Connecting Europe Facility

Supporting documents

- Dropbox with all relevant documents available through the following link: https://www.dropbox.com/sh/swp9naf090hpzwa/AACDFD33cEC6cV_e1CYCJN4Qa?dl=0
- Available information:
  - template for the letter of support to the national ministries and public service administrations (in English - to be adapted to your local context),
  - subcontract to support data collection efforts on your side (up to 4.500 EUR, also attached to this email),
  - the ELRC Resource Collection Guidelines (general process of data collection, questions regarding data sought, introduction to the ELRC-SHARE Repository
  - the overview of data collection activities in all countries (updated on a weekly basis)

European Language Resource Coordination
Connecting Europe Facility

Your ELRC representatives

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

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

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

- **DFKI** (Andrea Lösch – andrea.loesch@dfki.de, Lilli Smal – lilli.smal@dfki.de):
  - Germany
  - Austria
  - Luxemburg
  - Netherlands
  - Hungary
  - Czech Republic

European Language Resource Coordination
Connecting Europe Facility

Financial support to data collection I1

- Subject of the subcontract:
  - The overall goal of this subcontract is to ensure provision of language resources from <Country> that are relevant to ELRC. The Subcontract covers the delivery of up to six such language resources by the Subcontractor.
  - The delivery of such language resources typically involves the follow-up with workshop participants according to ELRC's Resource Collection Guidelines (Important note: coordination with the ELRC representative is compulsory; the ELRC representatives are listed in Annex I to this Subcontract).
- Payment:
  - The Contractor shall pay to the Subcontractor 750 EUR per language resource delivered and accepted (see Section 2 for requirements). The maximum amount is 4.500 EUR.

European Language Resource Coordination
Connecting Europe Facility

Financial support to data collection I2

- Current requirements for data sets:
  - Relevance:
    - The language resources must be produced by or relevant to public service administrations and public bodies in the Subcontractor's country.
    - The data sets should be relevant at least to the general (administrative/regulatory) domain, but ideally to the domains covered by the Digital Service Infrastructures (CEF DSIs). They could also significantly be resources that significantly expand the coverage of machine translation.
  - Types of data: The data sought by ELRC can take different shapes, including in particular:
    - aligned parallel corpora,
    - translation memories,
    - translation/language models,
    - comparable corpora.
  - Further requirements:
    - Expected size of the data set: 100.000 words or more
    - Uniform encoding, ideally UTF-8
    - Good alignment quality