Major objectives of the ELRC

The European Language Resource Coordination Consortium (ELRC)

A service contract for the EC

Khalid CHOUKRI
ELRA/ELDA
choukri@elda.org

On behalf of the CEF ELRC team
ELDA, ILSP, TILDE, DFKI, TAUS

Key messages on the specifics of the ELRC work

…..➔Now that Josef did set the scene …

- What are specifically our major goals
- Instruments and approaches to implement them … Timeline
  - Preparing the ground (<M6)
  - Data collections (<M24)

- Why we can not achieve them without You
- If we do so, how can we service the EU and our community (R&D, Industry, MS)

- How can we sustain the process and the efforts

Our major goals

- **Golden Goal … LR identification and Collection**
  - Secure 200+ Language resources suitable for use within Assisted Translation (AT//MT), and sustain the process

- **Instruments …. Awareness and Info/data sharing**
  - Establish and staff a helpdesk
  - Connect with national bodies and prepare the ground
  - Organize one seminar per country (30+), *improve readiness and ability to contribute data*
  - Establish a pipeline to collect data
  - Work out jointly for a win-win deal for all parties

Preparing the ground
Helpdesk and Support

Set up and run a technical-legal helpdesk that will help with all queries regarding language resource identification, preparation, processing and sharing.

- Technical aspects, such as formatting, encoding, metadata usage, metadata conversion to LOD, packaging, uploading, maintenance;
- Basic data processing, such as data cleaning, alignment, annotation schemas, data validation, processing evaluation,
- Legal aspects towards data sharing, comprising from licensing models and IPR clearing to data anonymization and confidentiality
- Administrative issues.

Preparing the ground the ELRC Website

- Key information about the CEF and ELRS in all CEF languages
- Overview, materials, findings, registration, calendar of the ELRC events - Workshops and Conferences
- Latest news and activities related to the ELRC
- Support for social media Facebook, Twitter and LinkedIn
- Feedback gathering facility
- Discussion forum
- Connect to the Helpdesk and FAQ for ELRC related issues
- Language resource repository facility – process, upload, access
- Document repository

Preparing the Ground
Country-specific training workshops

- **Objective:** Identification and evangelization of "decision" makers about
  - Multilinguality,
  - Useful role and support of MT, and
  - Requirements of current MT technologies in terms of data

- **How**:
  - Organize circa 30 training Workshops in EU Member and CEF-Affiliated States;
  - Organization in cooperation with local partners to seek a multiplier effect (e.g. national anchor points but also DGT local branches)
  - Provide a high level of localisation and adaptation of material; local speakers,

- **Targeted audience**
  - Mostly Decision makers in national « publications » offices and DSI-Like Mangers
  - Producers & right-owners of (Public Sector) and similar MT valuable resources

Preparing the Ground
Country-specific training workshops

➢ Planned/Expected outcomes:

- Identification and reaching local stakeholders … Opening the doors
- Boosting data awareness inc. PSI directive, logistic/legal aspects
- Emphasize the benefit of data providers (better AT/MT for My language)
- Identification of (usable) Data collections + (right)holders;
- Local support of a concerted pan- European Action orchestrated by ELRC on MT/AT; AT.DSI as a web-service /APIs.
- Identification of digital services that can adopt (early adopters) MT/AT technologies on the local scenes

(tentative) Organization of the schedule of workshops

- **The workshops will proceed in parallel:**
  - Initial schedule of the workshops in a "**rehearsal/initial phase**" (3-4 Workshops)
  - These workshops will serve as pilots for the second and broader phase.
    - Analysis of the outcome and re-tuning of the Workshop material & approach
    - Running the second round of the workshops (26+)

- **Geographic areas and respective responsibilities …**
  - Area 1 would be under the responsibility of Tilde (Latvia, Lithuania, Estonia, Finland, Sweden, Norway, Denmark, Iceland),
  - Area 2 under the responsibility of ELDA (France, Spain, Portugal, Italy, Malta, Belgium, the Netherlands, Luxembourg, Germany, United Kingdom , Ireland)
  - Area 3 under the responsibility of ILSP (Greece, Cyprus, Bulgaria, Romania, Croatia, Slovenia, Austria, Czech Republic, Slovakia, Hungary, Poland).
  - With the logistic support of DFKI and TAUS

The Success of Workshops

- Run all workshops with the next 6 months !
- The attractiveness of the workshops
  - (how many attendees, how many “Key” players, LR right holders)
- Good “feedback” from the surveys
- How efficient are the outcomes in terms of LR sets identified and secured
- Agreement in principle on LRs donation
- Adoption/Deployment of AT/MT within the national offices
- Report at the second ELRC Conference (likely @LREC 2016)

Collect LR Data Sets

- Identify and collect data sets, based on preparatory action, stakeholder and data leads in previous task;

- Target 200+ new data sets;
  - Alignment with EC on required data properties: domain, type and quality;
  - Set-up LR collection team (both from the consortium and identified contacts),
  - Collect Data:
    - » Identify and prioritize the sources of the data (online versus offline)
    - » Use traditional approaches to obtain data (partners expertise)
    - » Use new approaches i.e. for online sources will be crawled (tools from Panacea, QTlaunchPad, TaaS, etc.)
    - » Best Crawlers that allow to identify parallel data and comparable data in specific domains of knowledge will be used.

  - Perform basic data-cleaning, pre-processing, formatting, conversion and alignment where required; (automatic) quality review and maintenance;

Collect LR Data Sets

- Add some (basic) documentation and the necessary meta-data capture;
- Ensure clearing legal issues (IPR, Licensing, etc. where necessary)
- Set up and run a rigorous data quality control system using automated tools (based on ELRA quality Control methodologies) and manual sampling spot-checks
- Set up shared reporting tools
- Recoding and tracking progress against target;
- ***Set up and operating Storage and Distribution mechanisms (Meta-share)***
- Strong Partnership with the EU Open Data Portal (a “sharing” channel)
- Regular and Final delivery of data to EC.

Data Quality and Validation

- Data Quality vs meta-data quality, documentation, …
- Identify Sources of data (quality tag!)
  - some are reliable (official administration vs blogs from public officers)
  - Automatically assess the confidence in “genre”, “domain”, language register,
- (Automatically) Identify parallel/comparable data “levels”
- Define and Assess quality of data & meta-data
- Assess the “openness” of the associated licenses;
- Etc.

Data special ISsuES

- Data is not in digital format, not a known format ! (……) shapes and forms
- We have translated texts but the sources are lost
- No one knows what /who are right holders
- Data is text, lists , mono bilingual ….but what can we donate !!
- We only have PDF, OCRed data, Wseb-pages in HTML, Word documents, PDF documents, Excel sheets …
- May be even better translated texts, translation memories, etc.
  - have you ever had text translated by an external company? Who are they? Who owns what ?
  - What about personal information in the data? ELRC will “remove” this
  - and many other things …
- Unexpected issues
- PSI is a directive …. Enforcement but not our way of operating
- Other Legal issues , Ethics, etc.

Data special issues

- Data is not in digital format, not a k …) shapes and forms
- We have translated texts but th
- No one knows what /who ar
- Data is text, lists , mono l iate !!
- We only have PDF, O ML, Word documents, PDF documents, E
- May be even bett emories, etc.
  - have you external company? Who are they? W
  - Wh the data? ELRC will “remove” this
  -
- I
- cement but not our way of operating
- Ot nics, etc.

Success of Data Collection task

- How Many new resources have been identified /Quarter ?
- How Many new resources have been secured for use within AT.DSI ?
- How many are made widely available ?
- The MT@EC deployment at local administration , API/Web services for us

Concluding message: failure is not an option

Joint effort to support EU and MS but also our languages and hence our innovative players

Timeline of the action ….

Two Major phases

- Setting the ground (<M6)
  - Identify Contacts & Connections
  - Helpdesk
  - Workshops
- LR Collection phase (--M24)

- Information dissemination … Web , conferences, Social networks
- Urgent actions
  - Tune the messages for our contacts & stackholders ….
  - Seminars before summer
  - Rehearsal …. With your input involvement

The official tasks list …

- Task 1: Secretariat of the Language Resource Coordination (DFKI)
- Task 2: Technical Helpdesk for Language Resource provision (ELDA)
- Task 3: Language Resource Board (DFKI)
- Task 4: Website (Tilde)
- Task 5: Conferences (DFKI)
- Task 6: Targeted country-specific training workshops (ELDA)
- Task 7: Language Resource data sets (ELDA)
- Task 8: Advisory and consultancy services (DFKI)

Psi .....

- Amended Directive
- The main changes in the amended Directive are to:
- • require public sector bodies (PSBs) to allow the re-use of existing and generally accessible information they create, collect or hold. The effect of this was to make re-use mandatory in most cases.
- • extend its scope to cover PSI held by public sector museums, libraries (including university libraries) and archives in making their information available for re-use.
- • introduce the general principle that charges for re-use should normally be set at marginal cost, with exceptions in certain circumstances.
- • introduce a redress mechanism for complaints by re-users operated by an impartial review body with the power to make binding decisions