European Language Resource Coordination (ELRC)

Public Service Data for the Automated Translation DSI

Josef van Genabith. Khalid Choukri,
Andrejs Vasiljevs, Stelios Piperidis,
Jaap van der Meer

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

ELRC: Who are we?

- A tender action (2015-2016) under CEF AT.DSI

  - DFKI — Josef van Genabith
  - ELRA/ELDA — Khalid Choukri
  - TILDE — Andrejs Vasiljevs
  - ILSP Athena — Stelios Piperidis
  - TAUS — Jaap van der Meer

ELRC: Context

- Europe is multilingual
- Equal opportunities for all languages
- No discrimination, no silos
  - Digital Single Market = multilingual DSM
  - DSM = mDSM
  - Translation = enabler
- Technology support: Automated Translation DSI under CEF

ELRC: Context

- Statistical machine translation (SMT, Moses)
- All about data
- ELRC support CEF collect data for AT.DSI

ELRC: Context

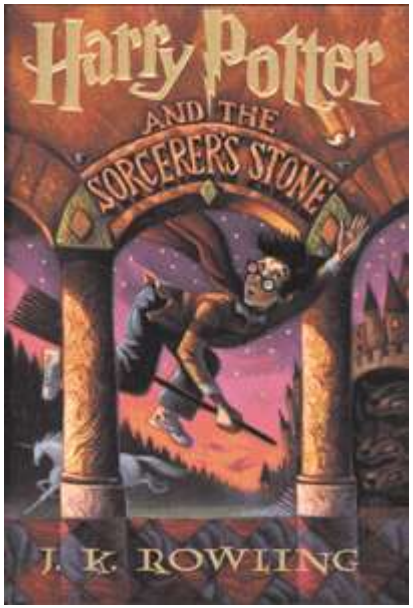
- Goal mSDM:
- Free flow of people, information, services, culture, and commerce

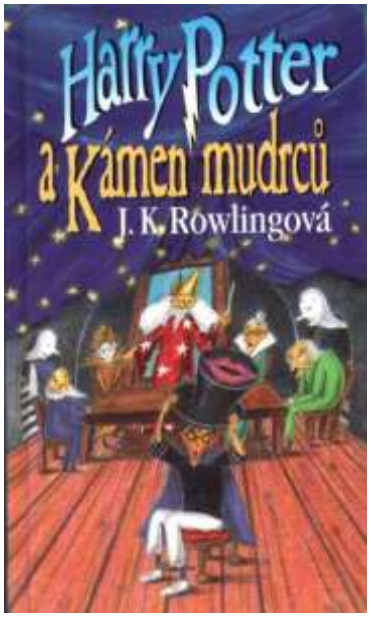
- Goal AT.DSI:
- national governments, public administration, public services, NGOs

ELRC: Data

- Statistical Machine Translation is all about data
- SMT learns how to translate from data
- Data
  - translations (bilingual data)
  - Monolingual data (target language text)
  - Dictionaries, terminology, ontologies, named entities
- Like people SMT is good at what it has learned

ELRC: Data

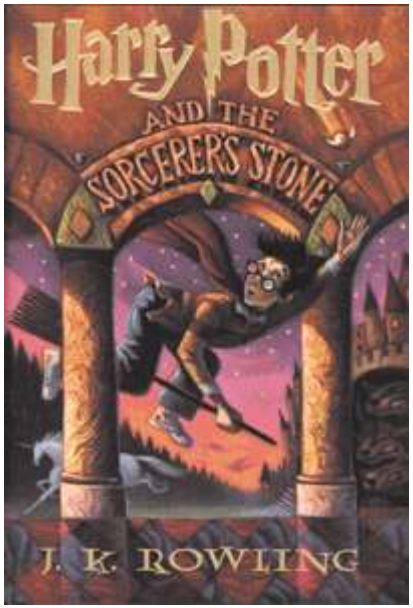


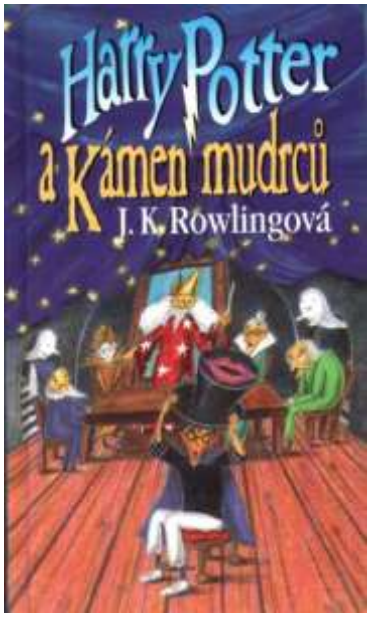


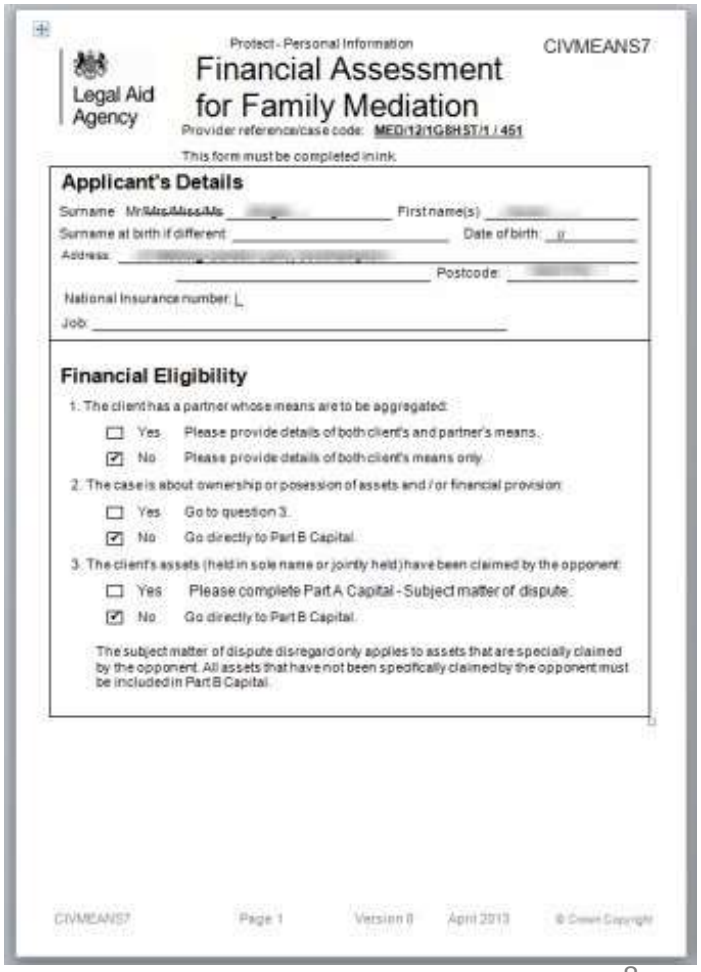

"Good," said Wizard Wink. "We'll start with Numbers."
"Oh no!" groaned Max and Mick.
The wizard boys didn't like school much. They wanted to play.
Wizard Wink asked them the first question.
"If I had three crystal balls in one hand . . ." he said, "and four in the other . . . what would I have?"
"BIG HANDS!" shouted Max and Mick.
"Very funny," said Wizard Wink.
"Oooh! I know. I know!" said Zoe. "Seven!"
"Right!" said Wizard Wink.
"Poo!" said Melissa. "That was easy-peasy."

MOSES CORE

ELRC: Data







Protect - Personal Information CIVMEANS7

Legal Aid Agency

Financial Assessment for Family Mediation

Provider reference/case code: MED/12/1G8HST/1 / 451

This form must be completed in ink.

Applicant's Details

Surname: Mr/Mrs/Miss/Ms First name(s):

Surname at birth if different: Date of birth:

Address: Postcode:

National Insurance number:

Job:

Financial Eligibility

1. The client has a partner whose means are to be aggregated:
   - ☐ Yes Please provide details of both client's and partner's means.
   - ☑ No Please provide details of both client's means only.
2. The case is about ownership or posession of assets and / or financial provision:
   - ☐ Yes Go to question 3.
   - ☑ No Go directly to Part B Capital.
3. The client's assets (held in sole name or jointly held) have been claimed by the opponent:
   - ☐ Yes Please complete Part A Capital - Subject matter of dispute.
   - ☑ No Go directly to Part B Capital.

The subject matter of dispute disregard only applies to assets that are specially claimed by the opponent. All assets that have not been specifically claimed by the opponent must be included in Part B Capital.

CIVMEANS7 Page 1 Version 8 April 2013 © Crown Copyright

MOSES CORE

ELRC: Data

- Need the right kind of data
- national governments, public administration, public services, NGOs
- CEF provide services for multilingual engagement with national citizens, EU citizens and other customers of public administration

ELRC: Data

- Data formats?
  - Word and text processing documents
  - PPT and other presentation software presentations
  - Spread sheets
  - PDF documents
  - Data bases
  - Content management systems
  - Web sites
  - Translation memories

ELRC: Data

- Where do we find this data?
  - National stakeholder institutions
  - Government
  - Administration
  - Public services
  - NGOs
  - LSPs

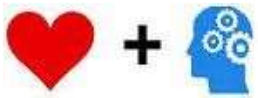
ELRC: Data

- Assumptions
  - Many public services hold relevant data
  - They don’t know this
  - If they know, they are mostly nor aware of the value of the data
  - If they are, they may be reluctant to provide the data

ELRC: Data

- Finding the right data
- Finding the right people
- Promote data awareness
- Help them find the data themselves
- Convincing them to provide the data (legal, privacy, services)

- How do we do this?
- Winning hearts and minds



ELRC: Data

- 24 official languages
- 28 member states
- 30 CEF affiliated countries = 28 + Iceland and Norway

ELRC: Local Ownership and Responsibility

- Local ownership and responsibility
- Local experts know best
- National Anchor Points
- Twins: 30+30
  - Highly renowned language expert
  - Representative of local government
- Conduits to national stakeholders
- + key reps of CLARIN, FlaReNet, LTI, …

ELRC: Language Resources Board

- Bulgaria Svetla Koeva
- Croatia Marko Tadic
- Czech Republic Jan Hajic
- Denmark Sabine Kirchmeier-Andersen
- The Netherlands Jan Odijk
- Belgium Walter Daelemans
- Estonia Kadri Vider
- Finland Krister Linden
- France Joseph Mariani
- Germany Andreas Witt
- Austria Gerhard Budin
- Greece Maria Gavrilidou

ELRC: Language Resources Board

- Hungary Tamás Váradi
- Iceland Eirikur Rögnvaldsson
- Ireland Andy Way
- Italy Simonetta Montemagni
- Poland Adam Przepiórkowski
- Portugal Antonio Branco
- Romania Dan Tufis
- Slovenia Marko Grobelnik
- Spain Núria Bel
- Sweden Lars Borin
- Norway Gard Jenssen, Torbjørg Breivik

ELRC: Language Resources Board

- NAPs already provided 302 National Institutions across Europe

- Two tasks for NAPs:

- Help us find the government part of the twin
  - Please suggest 2 or 3 with short background info
  - Appointment by EC
- Help us extend the 302 lead contacts to 600
  - Focus on public services

ELRC

- Help us make AT.DSI a success
  - Services for Europe’s citizens
  - Support multilinguality
  - Visibility for LT
- Supporting your/our language is supporting Europe