ELRC - Spain

Núria Bel

Universitat Pompeu Fabra

GOBIERNO DE ESPAÑA MINISTERIO DE INDUSTRIA, ENERGÍA Y TURISMO

(ReTeLe)

upf. Universitat Pompeu Fabra Barcelona

ELRC Workshop

- Was held in Madrid, 26-01-2016
- Supported by the Secretary of Telecommunications and for the Information Society (SETSI)
- Mr. Víctor Calvo-Sotelo, State Secretary supported the event.
- 81 Participants
  - Public Administration: 50
  - University: 19
  - Industry: 6
  - Other: 6

SETSI Actions

- Within the framework of the Plan de Impulso de las Tecnologías del Lenguaje (PITL) granted **ReTeLe Network of Excellence** a contract for producing an inventary of language resources for MT in Spanish Public Administration Organizations
  - Identification
  - Metadata completion & annotation

Finding Potential data holders / Providers

- Two profiles:
  - Organisms that outsource translation. Finding information about outsourcing in Public Procurement web portal.
  - Organisms with In-house translation services. In two steps:
    - Letter by SETSI reminding key concepts (Open Data) and collaboration request by filling a form
    - Personal interview to identify resources and annotating metadata: finding conditions of the dataset

Translation Contracts in Public Administration - outsourcing

- Study on data available at Public Procurement Web (difficult to assess volumes) for 2011-2016
- Most common:
  - Police
  - Justice
  - Tourism
  - Social Security
- Frequently for both interpretation and translation services (or other linguistic services)

Organisms outsourcing translation:

- Also for public web pages (3/59 approx.)
- Also for Machine Translation services (1/59, licences)

In-house Translation, departing from:

- White paper on Institutional Translation in Spain (2011)
  http://ec.europa.eu/spain/pdf/libro_blanco_traduccion_es.pdf

From 136 questionary answers.
Resources available at your work place

-Paper dictionaries: 75.5%
- Computer: 85%
- e-Dictionaries: 23.5%
- Limited internet access: 57.7%
- No internet access: 31.6%
- **CAT tools: 7.35%**

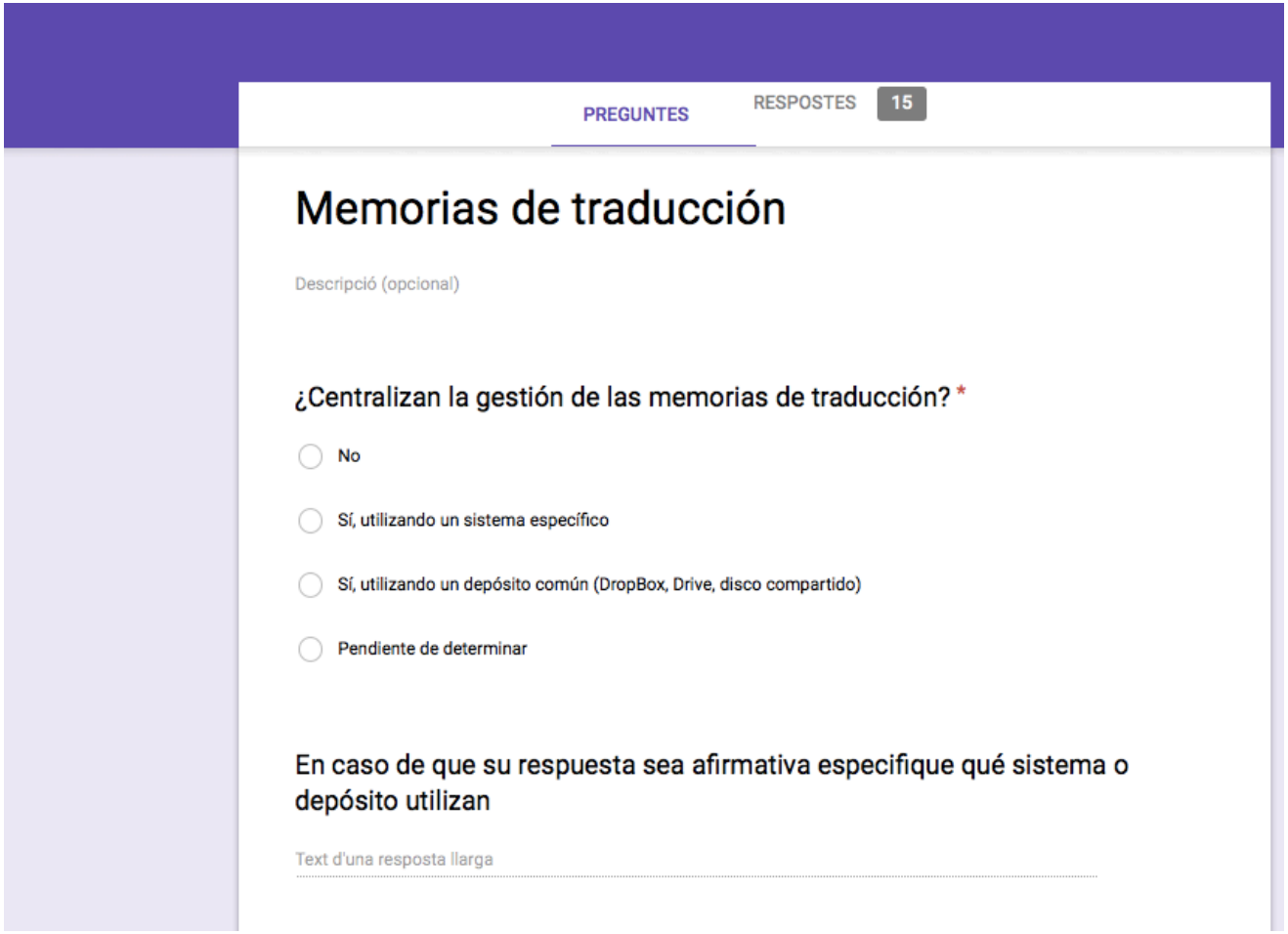
In-house Translation

- Contact ELRC workshop participants for a survey to identify actual resources/providers:
  - 24 organizations contacted (May – June)
  - 10 organizations responded (as for 6-07-16)

Personal interview

- Find out factual data about resources
  - Finding out technical details
  - Finding out legal issues
  - Identifying those for Open Data Portal
- Prioritization by form answers: identify larger providers

The form, a support to classify following the Maturity Model



The survey: Form follows …

Organizations									
LEVEL	Archiving	Document × file × language	PDF	Plain text .txt, .odt, .html, .docx	Aligned documents	Translation memory × document: sentence - aligned	TMX	Translation Memory x domain / areas	Standard Metadata
0		✓	✓						
1		✓		✓					
2	✓	✓		✓	✓				
3	✓	✓		✓	✓	✓			
4	✓	✓		✓	✓	✓	✓	✓	
5	✓	✓		✓	✓	✓	✓	✓	✓

Findings

- Findings (by now):
  - 3/10 works with CAT
    - 2 centralize archive of TMs
    - 1 using TMX
    - 0 protocols for updating TMs
  - 6/10 have centralized archive of translations & docs, in text processor format
  - 4/10 have documents more than 300 words per document
  - 3: 100, 1000 and more than 50,000 documents
  - 4/10 confidential data

Level	MATURITY CHARACTERISTICS	Risk & Cost
0	PDF	Difficult to predict the results of conversion
1	No defined archiving process, no document-language-ID easy identification	Inconsistency, low q of resources
2	Common Archive managed by a protocol, document-language-ID alignment	No quality control. Extraction and alignm still required.
3	Translation memories are integrated in the translation process and standard procedures are defined and are part of staff training	Gathering of data and publication right clear still required
4	Translation memories stored and managed centrally as an internal resource, but manually documented and updated by translators	Licensing schema still required
5	Translation memories are considered a licensed good to be shared. Documented with standard metadata, the protocol contemplates publication	

4

3

3

analysis to come ...

LEVEL	languages	respon-sable	size in seg-men-ts / words	creation data	domain	character encoding	associated resources	documentation	private data	confidenciality	licence
0											
1											
2	✓	✓									
3	✓	✓	✓	✓							
4	✓	✓	✓	✓	✓	✓	✓	✓			
5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

How to fit with ELRC requirements

- Characteristics of a resource
  - 100.000 words (about 350 pages)
  - utf8
  - good aligment

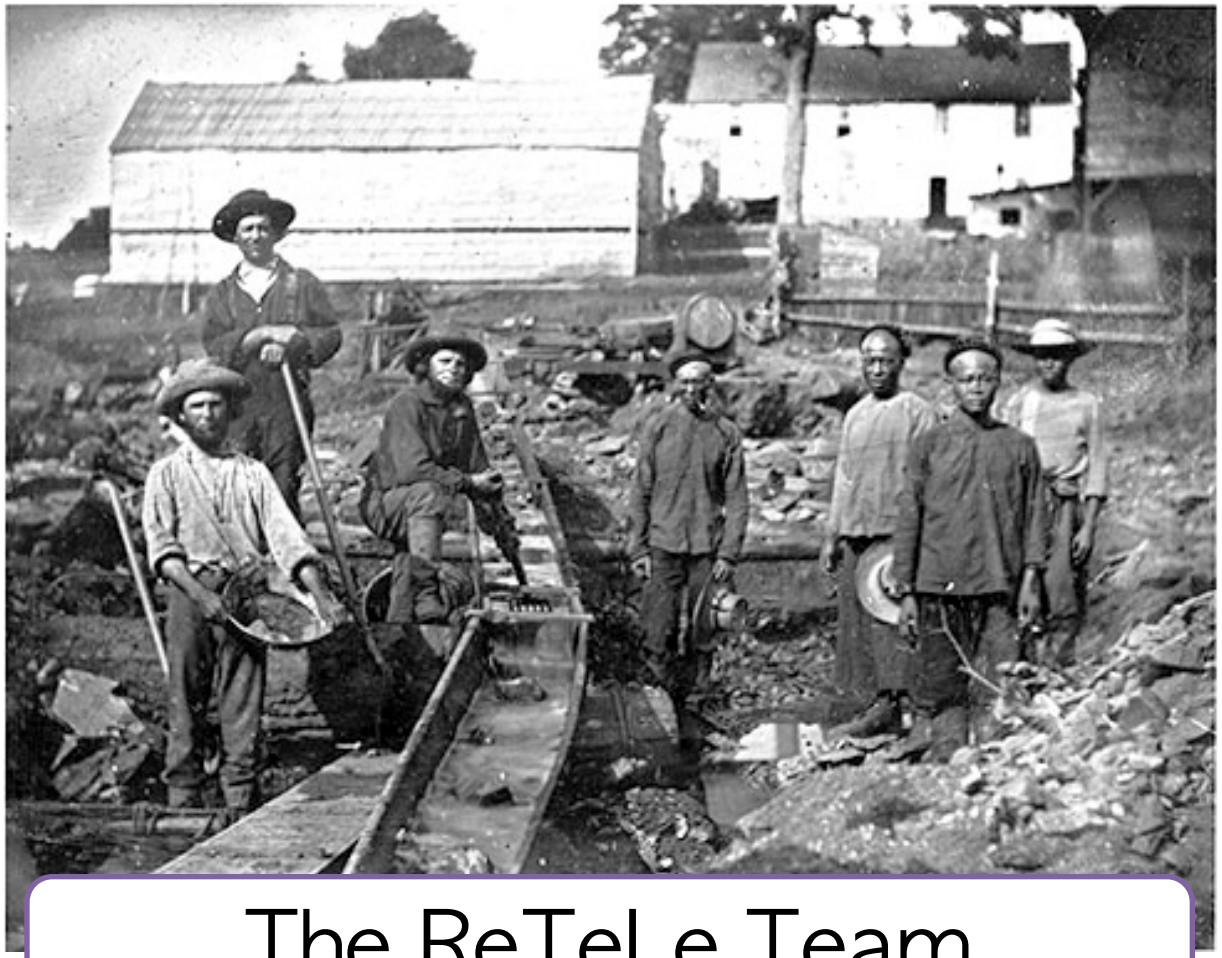
(Early) Conclusions

- (This is going to be hard!)
- 3 web sites can be crawled/request texts, expected hight quality translation (Tourism and Social Security)
- Alliance with DTIC to provide support for translators (CAT and MT) if agree to deliver TM.

Next actions

- Immediate actions: personal interviews
- Workshop for user requirements Translation services: CAT and MT and delivery of TM: anonimization, central repository, protocols ...
- Report writing

Thanks!



The ReTeLe Team

GOBIERNO DE ESPAÑA
MINISTERIO DE INDUSTRIA, ENERGÍA Y TURISMO

(ReTeLe)

Universitat Pompeu Fabra Barcelona