META FORUM 2016

A Maturity Model for Public Administration Open Translation Data Providers

Núria Bel (Universitat Pompeu Fabra)
Mikel Forcada (Universitat d'Alacant)
Asunción Gómez-Pérez (Universidad Politécnica de Madrid)

ReTeLe Network of Excellence
http://retele.linkeddata.es/ (TIN2015-68955-REDT)

upf. Universitat Pompeu Fabra Barcelona

Mining for Public AdministrationTranslation Data

Two recent initiatives:

- ELRC for CEF.AT

  http://www.lr-coordination.eu/

- Plan Nacional de Impulso de las Tecnologías del Lenguaje

  http://www.agendadigital.gob.es/planes-actuaciones/Paginas/plan-impulso-tecnologias-lenguaje.aspx

The mission:
Find reusable Translation Data!

- **Translation memories**
- **Corpora**
- **Lexica & Terminological resources**

Feasibility, a clue:
White paper on Institutional Translation in Spain (2011)

3. Medios de que dispone para su trabajo

—Diccionarios en papel: **103**
—Ordenador
Propio: **83** Compartido: **33**
No: **20**
—Diccionarios en formato electrónico: **32**
—Acceso a Internet:
Limitado: **78** Ilimitado: **15**
No: **43**
—Programas de traducción asistida: **10**

Resources available at your work place?
(From 136 answers – 7 org.)
- Paper dictionaries: 75.5%
- Computer: 85%
- e-Dictionaries: 23.5%
- No internet access: 31.6%
- **CAT tools: 7.35%**

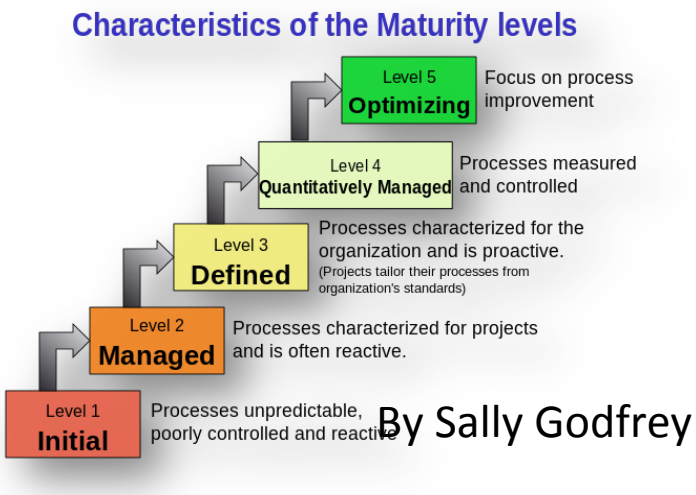
http://ec.europa.eu/spain/pdf/libro_blanco_traduccion_es.pdf

Assessing the feasibility of 'the mission'

- Advantages
  - Producers of lots of data of all domains
  - Policies about public information reusability
- Problems
  - How many organisms in Public Administration are aware they are producing *translation data*?
  - What are the costs of identifying and packing these data as reusable data? Formats, archiving, licenses ...
- A proposal for supporting organizations with a Maturity Model: the path to optimal supply.

A Maturity Model for organizations to become providers of reusable Translation Data

These organizations should become aware of the goods they are producing and that adapting their processes to become optimal providers is also a path for getting better working conditions.



By Sally Godfrey

Maturity Model. Optimal?
Open Data and EU Reuse Policy

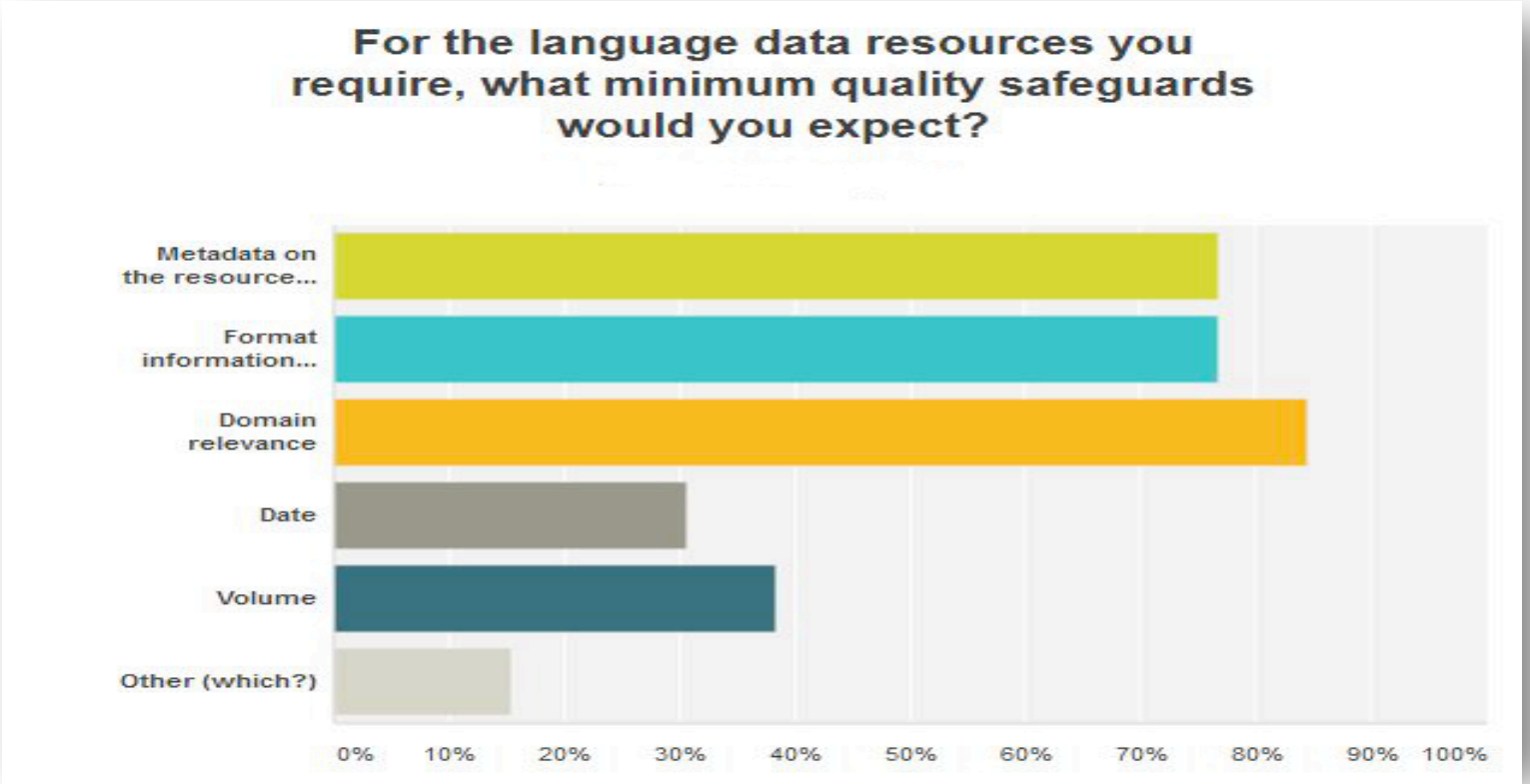
The full Open Definition gives precise details as to what this means. To summarize the most important:

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Re-use and Redistribution:** the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

Maturity Model. Optimal? Expectations about quality of data

- LT-Observatory survey

http://ltinnovate.blogspot.com.es/2016/04/the-future-of-language-resources-for.html
Resources catalogue: http://www.lt-innovate.org/lt-observe/resources-list



Maturity Model takes into account the cost of ...

- **Size:** e.g. cost of collecting many small documents?
- **File Format**: only processable text
- **Alignment**: from document-language identification to sentence/segment alignment
- **Publication limitations**: copyright, private data, confidential data
- **Cheap metadata annotation**: languages, size, domain, character encoding, license of the resource and clerical information: creation date, contact data of the responsible person, whether there is associated documentation and resources

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 quality of resources
2	Common archive managed by a protocol, document-language-ID alignment	No quality control. Extraction and alignment still required.
3	Translation memories are integrated in the individual translation process. Archiving standard procedures are defined and are part of staff training	Gathering of data and publication right clearance 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	

Maturity Model for organizations: Summary

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	✓	✓		✓	✓	✓	✓	✓	✓

Standard Metadata Summary

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	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Thanks!

- Please, this is ongoing work, comment, criticize and suggest ... at Poster time!