UKRAINIAN TRANSLATION INDUSTRY CONFERENCE

UTIC

Webinars-2015

Translation and multilingualism in the era of Big Data and Open Data

Kimmo Rossi

December 10, 2015

Host

Valentyna Kozlova
CMO of the UTIC

Speaker

Kimmo Rossi
Head of Research
and Innovation sector
DG CONNECT
European Commission

- Kimmo Rossi graduated in 1989 from Helsinki University of Technology (physics, IT, economics).
- Until 1994 he was co-owner and board member responsible for technology, quality and contractual issues for a Finnish translation and localization company Trantex.
- In 1994 he joined the European Commission where he worked 5 years in the translation tools and terminology department.
- Since 2000 he has been managing science, technology and innovation programs in the fields of statistics, data & content management and language technologies.
- Currently, he leads the Research & Innovation sector of Data Value Chain Unit (CNECT.G.3) managing a portfolio of 150 technology projects.

UKRAINIAN TRANSLATION INDUSTRY CONFERENCE UTIC

Outline

- definitions of some key concepts
- how Big Data – Open Data – Linked data affects translation/localization
- value creation by rich translation data
- some relevant EU projects
- conclusions

What is Big Data (BD)?

Data is "big" if it defies traditional processing & storage paradigms – "bigness" becomes the main problem

The "3Vs":

- Volume (size),
- Velocity (bits/s),
- Variety (db, jpeg, video, numbers, text in language X...)

...more V's can be added:

- Veracity (provenance, quality, reliability...)
- Value (what BD can bring after "refining")

Linguistic Big Data

Most of the data on the web is human language – and many different languages

- Volume (size) – the Web as a corpus
- Velocity (bits/s) – e.g. 6000 tweets per second (≈ one book); over 30 million books a year (≈ Library of Congress)
- Variety (well, Twitter again: jargon, cryptic abbreviations, different languages)

...more V's can be added:

- Veracity – e.g. human (quality) translations vs (raw) machine translations
- Value – the topic of this presentation!

What is Open Data (OD)?

A useful definition: http://opendefinition.org/

- free of charge (or almost)
- free to use, re-use, distribute
- ...for any purpose (also commercially, also derivatives)
- can be government OD or private OD (user-generated, corporate)
- *must not violate privacy*

- www.open-data.europa.eu (EU Institution Data)
- http://www.europeandataportal.eu/ (EU Member State Data)
- **Public Data ≠ Free Public Data ≠ Open Data**

European Union Open Data Portal

Find datasets...

Show results with: ◉ all of these words | ○ any of these words | ○ the exact phrase ?

Total datasets available: 7851

Suggest a dataset

Is there a dataset from the EU that you could not find in this portal?

Please request the dataset>>

Most viewed datasets

view all »

- **DGT–Translation Memory**
  (13064 views)
- **Elevation map of Europe**
  (9895 views)
- **CORDIS – EU research projects under Horizon 2020 (2014–2020)**
  (9247 views)

Browse datasets by subject

- Employment and working conditions
- Social questions
- Economics
- Finance

What is Linked Data (LD/LOD)?

Allows access to Web content as a database – very useful!

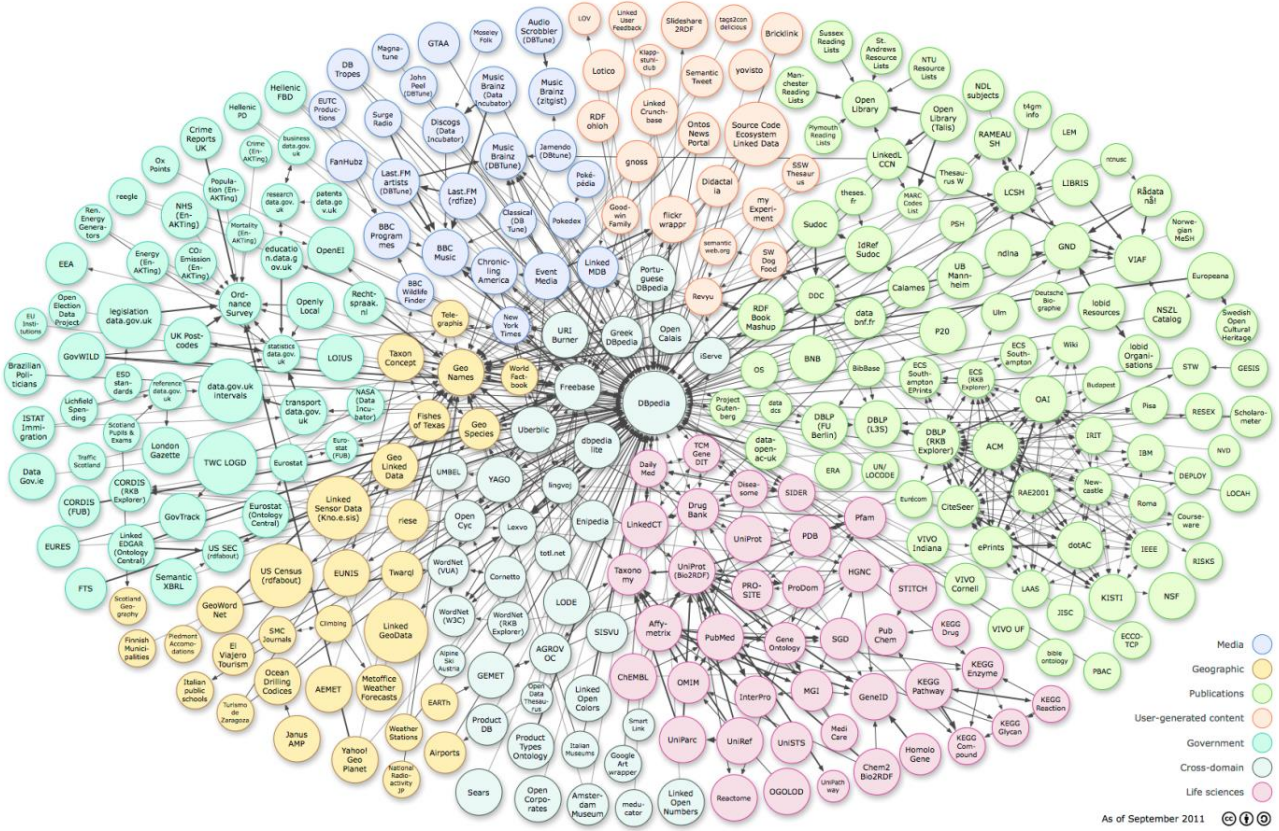
Semantic relations (meanings) in machine-readable form

But there are problems:

- What merits to be linked?
- How to establish links?
- How to ensure quality of (automatically established) links?
- Currently very scarce and English-dominated
- Example: statistics of EUROSTAT http://eurostat.linked-statistics.org/

More info: see LOD2 project: http://lod2.eu

Linked Open Data Cloud



Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

UKRAINIAN TRANSLATION INDUSTRY CONFERENCE UTIC

Linguistic Linked Data (LLD/LLOD)

WikiPedia is a "human-readable" equivalent of (linguistic) linked data

DBPedia is the machine-readable rendering of WikiPedia (but only a fraction of WikiPedia knowledge)

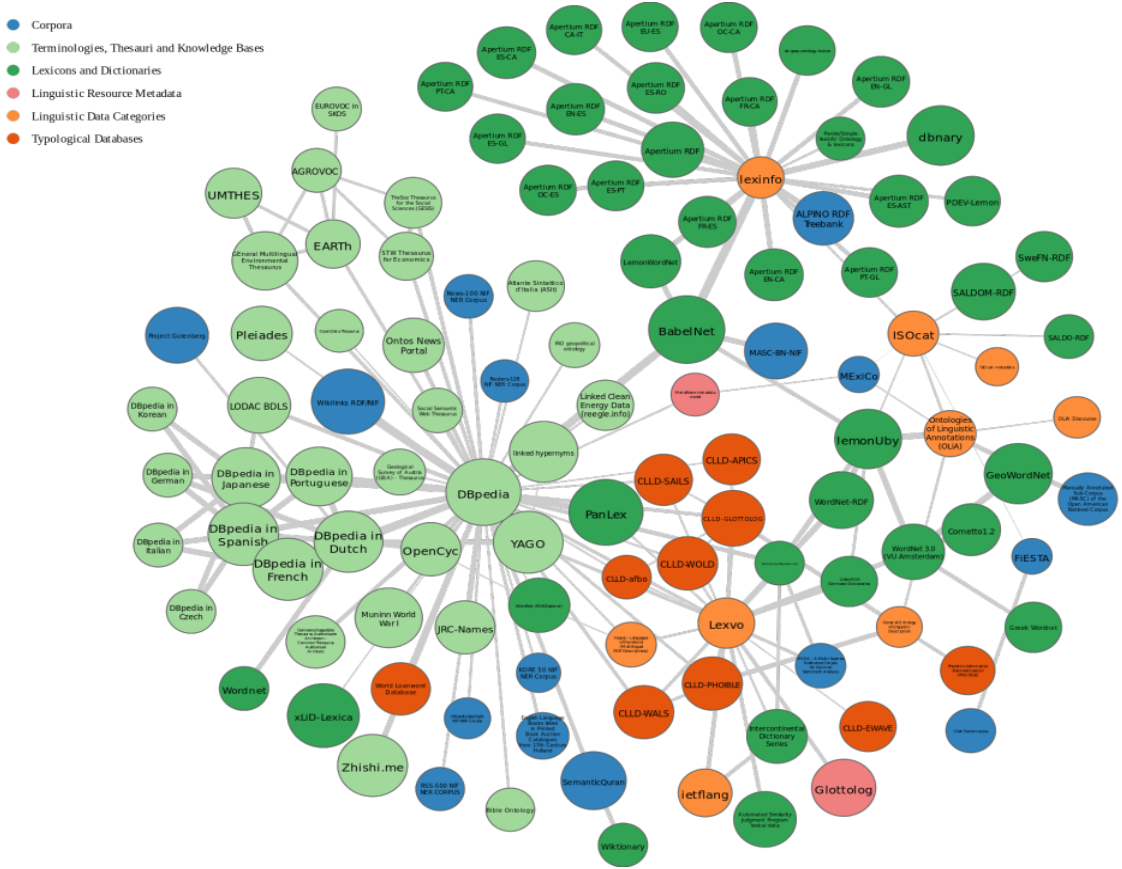
Raw linguistic data (e.g. documents, translations) is mostly unlinked

Linking adds value (e.g. from document pairs to sentence-aligned TBX file)

http://www.lider-project.eu/

http://www.linguistic-lod.org/

Linguistic Linked Data Cloud



Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

UKRAINIAN TRANSLATION INDUSTRY CONFERENCE UTIC

Challenges in translation business

Technology becomes more important

- integration of tools/platforms

Data becomes more important

- availability – accessibility - sharing

Processes & workflows are difficult & distributed

- MT, TM, SaaS, freelance networks, crowdsourcing

Prices/revenue go down – dumping of word prices

- selling just "words" (once) becomes less profitable
- better: sell services – processes – (quality) data, for re-use

Translation data

Legal issues

- who owns translation/TM/MT engine?
- contract becomes crucial – determines price
- micro-contracts, click transactions
- am I allowed to share data / harvest somebody else's data?
- confidential/personal data

High-quality data is valuable asset

- human translations – valuable raw material for re-use (MT/TM)
- add value by linking/annotating
- metadata adds value

Adding value with metadata and annotations

Source

Under **regional policy**, for example, the EU supports music schools, concert halls and recording studios and funds the restoration of historic theatres (e.g. the *Teatro del Liceu*, Barcelona, and the *Fenice*, Venice).

Translation

W ramach działań realizowanych w obszarze **polityki regionalnej** UE przyznaje środki szkołom muzycznym, filharmoniom i studiom nagraniowym, finansuje też renowację zabytkowych teatrów (np. barcelońskiego *Teatro del Liceu* czy weneckiej *La Fenice*).

Segment metadata:

Segment: Europa.eu-41595	Date: 03/01/2014	Link_f
Author: EAC.C.5_3303	Language: EN (UK)	Link_b

Translator:	AW-1222	Date: 15/05/2014	Link f
Language: PL		Revision status: 5A	Link b

File/document metadata:

Project: EW12555 TM file: EAC_36N6609 format: TBX_20 size: 2.304 languages: EN, DE, FR, PL, ES, IT

Licensing metadata (ODRL):

Assigners:	Offer:	Derive: ...	Agreement:	Privacy policy: ...
Permissions: ...	Prohibitions:....		Delivery_channel:....	Media:....

Projects: FALCON

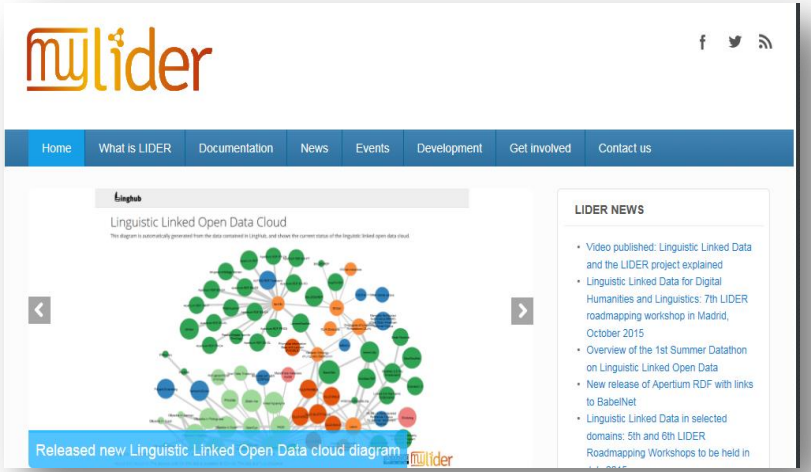
Tools for the Localization Web

- The Web as a database of language resources: web as corpus/TM/term database
- Tools for metadata creation and curation, using existing standards
- Demonstration of next-generation localization workflows
- http://falcon-project.eu/

Projects: LIDER

Works on *Linguistic Linked Data*

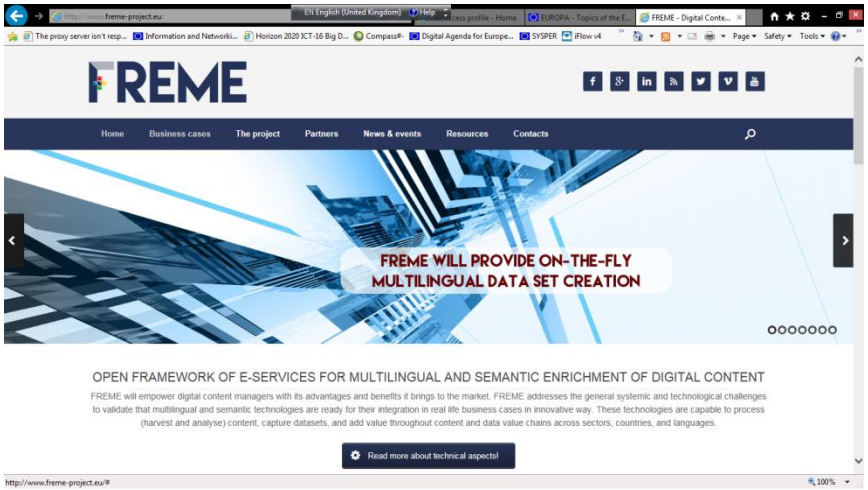
- Guidelines and best practices
- Community-building (W3C)
- Reference architecture for LLD
- http://www.lider-project.eu/



Projects: FREME

Semantic enrichment of multilingual content

- Automatic recognition of places, persons, events
- Retrieving additional information about them
- Converting human-readable (unstructured) into machine-readable (structured) representation
- 4 business cases (e-books, localization, food data, web recommendations)
- http://www.freme-project.eu/



Conclusions

Open translation data becomes "public good" – but not all translation data will ever become "Open"

Can a market emerge for translation data which is not Open?

- you can "sell" your translation 100 – 1000 – or million times!
- per-word price of translations go close to zero (or: to zero)
- tools for metadata generation, curation and using are necessary

- contracting/licensing needs to become "automatic"
- machine-readable metadata allows automatic re-use/trading
- metadata defines quality (=fitness for purpose)
- trading/re-use of confidential/personal data is difficult (but not impossible) – privacy-preserving technologies supported under Horizon 2020

Thank you!

kimmo.rossi@ec.europa.eu
lu.linkedin.com/in/kimmorossi/
Twitter: kimmorossi

UTIC videos and presentation materials

UTIC-2013/14 videos are published at
2015.utic.eu/video/

You can register to the forthcoming webinars and get familiar with videos of the previous at
webinar.utic.eu/

Join UTIC in the social networks!

for English-speaking audience:

- https://www.facebook.com/UTIConf
- https://plus.google.com/+UticEu
- https://twitter.com/UTIConf
- https://www.youtube.com/user/UTIConf
- https://www.linkedin.com/grp/home?gid=4698198

for Russian and Ukrainian-speaking audience:

- https://www.facebook.com/groups/UTIConf/
- http://vk.com/uticonf

Our guest for the next webinar

Tetyana Struk

CEO at the Linguistic Centre®

Translation and Localization company

December 24, 2015

UKRAINIAN TRANSLATION INDUSTRY CONFERENCE UTIC

UTIC-2016

Dnipropetrovsk, Ukraine

We thank UTIC-2015 partner