Multilingual Data Challenges in Professionalizing Data Stewardship worldwide

Romain David, Alison Specht, Margaret O'Brien, Lesley Wyborn, Christina Drummond, Rorie Edmunds, Claudia Filippone, Jeaneth Machicao, Nobuko Miyairi, Graham Parton, et al.

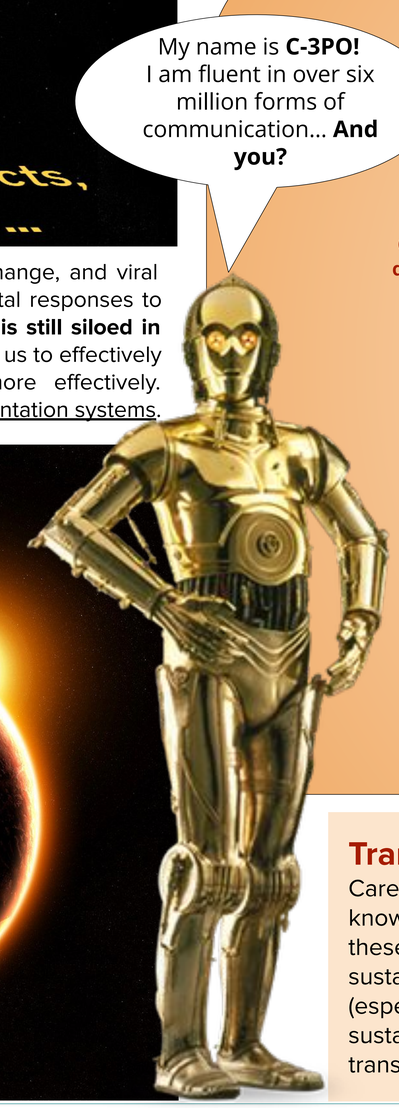
Multilingual Data Challenges in Professionalizing Data Stewardship worldwide

Authors: Romain David[1], Alison Specht[2], Margaret O'Brien[3], Lesley Wyborn[4], Christina Drummond[5], Rorie Edmunds[6], Claudia Filippone[1], Jeaneth Machicao[7], Nobuko Miyairi[8], Graham Parton[9], Debora Pignatari Drucker[10], Shelley Stall[11], Niklas Zimmer[12]

Contact: R. David romain.david@erinha.eu @Romain_DAVID_13 **Websites**: https://parsecproject.org/ https://www.erinha.eu/ **twitter**: @PARSEC_News @ERINHA_RI

The use of common FAIR vocabularies, that are **both human and machine readable**, is a key criterion in the FAIR principles (e.g. Principle I2 of Wilkinson et al 2016 specifies '(meta)data use vocabularies that follow FAIR principles'). Using common FAIR vocabularies will enable data interoperability and the necessary meta-analyses even when data have different origins and are based on multiple vocabularies. The objective of this poster is to offer **an overview of the many multi-language challenges for effective Data Stewardship**. For instance, some bottlenecks are highly dependent on community approval processes that are linked to data dictionary understandability, and/or related training challenges.

Very few, if any, of our present problems such as biodiversity decline, climate change, and viral pandemics stop at national, disciplinary and linguistic boundaries, yet our most vital responses to the shared problems, the information generated to analyze and derive solutions, **is still siloed in different languages and locations throughout the world.** It is clear that in order for us to effectively respond, we need to collaborate globally and communicate information more effectively. Globalization of research requires interoperability of our observations and experimentation systems.



Good practices

For translators:
- Scientific skills with at least good command of target languages
- Validators to check work translated
- A clear versioning system
- A translating strategy with prioritization levels

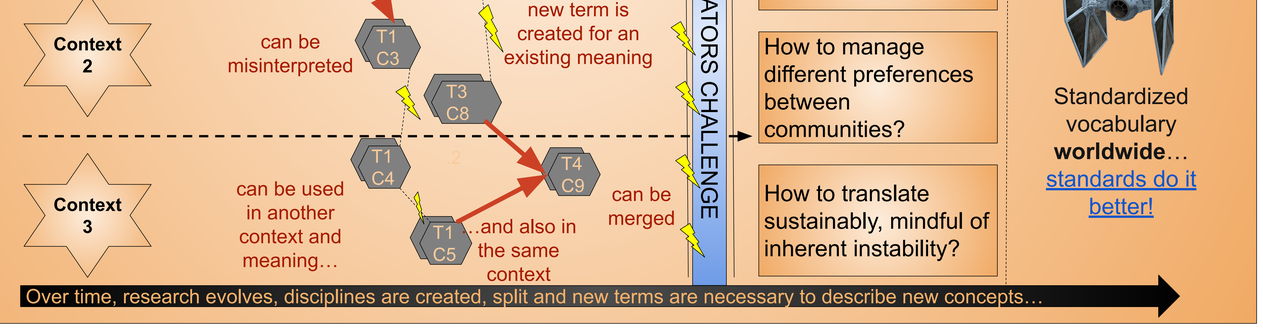
For Databases:
- Data dictionary with
  - disambiguated definitions both linked to and linked by concepts in shared ontologies
  - the date of the definition
  - the context of use (disciplinary - protocol - necessary skills...)
- A description of the community approval process (that can be adopted for new terms by other communities)

Science must be universally shared, in an interdisciplinary way, respecting that all disciplines are creating new concepts. The fact that science is always creating them is the core challenge for translators

The "TERM WARS" at global Scale: a time and space story

Discrepancies between regions and groups (culture, content, workflows, language, semantics, translation, funds...) are numerous whether for individual data users or more universally. We must anticipate issues such as **which is the preferred language**, **polysemy** (1 term, multiple meanings), **confusion** (multiple terms for 1 meaning or 'false friends' between 2 languages), plus **existing and evolving nuances** (not an exact match between languages and during time). Furthermore, terms are often **adopted from another language with different contexts and disciplinary realms** (that might decrease interoperability) and impedes **translation of all versions at the same time**.

Translation occurs at the concept level, *not as a simple one to one translation of (consecutive) words*.



Translation challenges are also organisational and sustainability challenges:

Care must be taken to ensure that datasets resulting from projects that practice co-creation and co-evolution of knowledge are translated into indigenous languages such that they can be used by the affected communities. Taking these challenges into account, we have to consider **human effort and the level of translation**, e.g. a low or minimum yet sustainable level, that is *legally allowable*. How can such minimal objectives be linked with FAIR principle compliance (especially FAIR and community approved vocabularies)? In several communities, translation is voluntary. One of the sustainability challenges is the need for ongoing engagement: how to keep interested groups involved. We need expert translators, as described, to **maintain the quality level critical to achieve effective harmonization** among languages.

Acknowledgements:
PARSEC is funded by the Belmont Forum through the National Science Foundation (NSF), The São Paulo Research Foundation (FAPESP), the French National Research Agency (ANR), and the Japan Science and Technology Agency (JST). This work is partially funded by the EOSC-Life European program (grant agreement No. 824087).We acknowledge wikipedia for R2D2, C-3PO and Tie fighter pictures.

Author affiliations : [1]ERINHA (European Research Infrastructure on Highly Pathogenic Agents) AISBL, FR, ORCID: 0000-0003-4073-7456; [2]The University of Queensland, AU, ORCID: 0000-0002-2623-0854; [3]University of California Santa Barbara, USA, ORCID:0000-0002-1693-8322; [4]Australian National University, AU, ORCID: 0000-0001-5976-4943; [5]Educopia Institute), ORCID: 0000-0001-5794-0413; [6]DataCite; [7]University of São Paulo, ORCID: 0000-0002-1202-0194, [8]Nobuko Miyairi, ORCID: 0000-0002-3229-5662, [9]Center for Environmental Data Analysis, ORCID: 0000-0003-4157-0352, [10]Embrapa Digital Agriculture, ORCID: 0000-0003-4177-1322, [11]American Geophysical Union, ORCID: 0000-0003-2926-8353, [12]University of Cape Town ORCID: 0000-0001-8078-0403

RDA 19th Plenary Meeting, Part Of International Data Week, 20–23 June 2022, Seoul, South Korea