LMF Revisited

Anas Fahad Khan[1], Francesca Frontini[1], Laurent Romary[2]
[1]Istituto di Linguistica Computazionale "A . Zampolli", CNR, Italy
[2]INRIA / Paris, France
[1]{*fahad.khan, francesca.frontini*}*@ilc.cnr.it* [2]*laurent.romary@inria.fr*
RELEVANT UNIDIVE GROUPS: WG2, WG3

Introduction

- Shared standards are essential in ensuring the interoperability and re-usability of language resources. This is especially important when it comes to efforts at promoting and maintaining language diversity via the such resources.
- It takes on an extra relevance in the case of computational lexicons contatining highly structured information which can be rendered explicit using standards such as the influential **Lexical Markup Framework (LMF)**, first published in 2008 by the International Standards Organization (ISO) as **ISO standard 24613:2008** and intended as a "standardized framework for the construction of computational lexicons" " (Francopoulo, 2013).
- In this poster we take a brief look at the original LMF and provide an update on te new version of LMF!

Meet the Old LMF (24613:2008) ....

- The original LMF specifications were intended to meet the need for a standard for lexical resources that would place a high priority on re-usability and interoperability. This was to facilitate a greater level of data exchange and to promote the merging and/or linking together of different individual resources and thereby avoid the proliferation of data silos.
- It is important to note that LMF was conceived of during a period of increasing recognition of the value of language resources for NLP, and of the importance of the re-usability and interoperability of data, something recently enshrined in the formulation and widespread adoption of the FAIR principles.
- The original LMF specifications were intended to cover as wide a range of lexicon-like resources as possible. Hence they made specific provision for both NLP dictionaries and *Machine Readable Dictionaries*, as well as several other categories of lexicon or lexico-semantic resource, such as for example *bilingual* and *multilingual lexicons* along with Wordnets. In addition, the original specifications were designed to take a wide range of linguistic information into account. In particular the original LMF specifications consisted of a core model together with the following series of extension packages: **Machine-readable Dictionary**, **Morphology, Syntax and Semantics**, **Multilingual Notation**, **Multiword Expression Pattern**, **Constraint Expression**.
- Special care was taken to ensure that the specifications were not exclusively 'euro-centric' and that non-European languages were very much taken into consideration during the drafting of the standard (see (Francopoulo, 2013)).

Not the same as the New LMF (ISO 24613-1-6)

- The original version of LMF allowed for data modelling at several different levels of linguistic description. This **led to significant complexity** in the resulting standard, something that was handled through organising the standard into **separate packages**.
- Users were obliged to consume the standard as a whole even if they were only interested in specific parts and several salient areas of linguistics/lexicographic description such as etymology were not covered at all in the original LMF
- The lack of modularisation/de-coupling made the (inevitable) addition of new parts awkward (especially given the ISO workflow for publishing materials)
- So the ISO sub-committee **ISO TC 37/SC 4/WG 4** decided to create a new version of the standard which would address these issues.
- The **new LMF** is a multi-part standard consisting of six separate modules, each published as a separate ISO standard, with further extensions planned to come.
- It has been decoupled from any single serialisation format, although two recommended serialisations of the meta-model constitute the fourth and fifth parts of the standard (these are TEI and LBX respectively).
- The new emphasis on abstraction and modularisation has also led to a series of major simplifications affecting nearly every part of the new version of the LMF meta-model. In the rest of this submission we list the new parts of LMF which have either been published or which are under development

ISO 24613-1:2019 Language resource management — Lexical markup framework (LMF) — Part 1: Core model: This module defines the basic classes required to model a baseline lexicon and is a pre-requisite for the use of the other classes. *Status: Published in 2019 it is now being further revised to make it easier to use.*

ISO 24613-2:2020
Language resource management — Lexical markup framework (LMF) — Part 2: Machine-readable dictionary (MRD) model
Contains components providing deeper specification of lexical description encapsulated within the core model. *Status: Published in 2020.*

ISO 24613-3:2021 Language resource management — Lexical markup framework (LMF) — Part 3: Etymological extension: A completely new addition to the LMF meta-model covering etymological and diachronic information. This part makes etymologies, etymological links and etymons first class citizens. See (khan2020towards) for more details. *Status: Published in 2021.*

ISO 24613-4:2021 Language resource management — Lexical markup framework (LMF) — Part 4: TEI serialization:
A TEI serialisation of the other parts of the model which aims to make both TEI and LMF fully compatible and which leverages the knowledge and makes use of the established practices of the TEI community in dealing with lexiocraphic resources. *Status: Published in 2021.*

ISO 24613-5:2022 Language resource management — Lexical markup framework (LMF) — Part 5: Lexical base exchange (LBX) serialization: Another XML serialisation. *Status: Published in 2022.*

ISO/CD 24613-6 Language resource management — Lexical markup framework (LMF) — Part 6: Syntax and Semantics}: An update to the Syntax and Semantics parts of the original standard. *Status: A candidate for an ISO Draft International Standard (DIS) ballot.*

Gil Francopoulo. 2013. LMF lexical markup framework. John Wiley & Sons
Fahad Khan and Jack Bowers. 2020. Towards a lexical standard for the representation of etymological data. *Convegno annuale dell'Associazione perl'Informatica Umanistica e la Cultura Digitale.*
Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Banski. 2019. Lmf reloaded. *arXiv preprint arXiv:1906.02136.*