LMF Revisited

Fahad Khan and **Francesca Frontini**
CNR-ILC / Pisa, Italy
firstname.secondname@ilc.cnr.it

Laurent Romary
INRIA / Paris, France
laurent.romary@inria.fr

Relevant UniDive working groups: WG2, WG3

1 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 which tend to contain highly structured information which can be rendered explicit using markup languages and where standards can help ensure a shared semantics. One such standard is 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 the current work we take a brief look at the original LMF and explain why the decision was made to update it as a multipart standard. We also provide an update on this new version of LMF, following on from that given in (Romary et al., 2019).

2 LMF - The 2008 Version

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[1], 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)).

3 The New LMF

The original version of LMF allowed for data modelling at several different levels of linguistic description. Understandably, this led to a significant amount of complexity in the resulting standard, something that was handled through the organisation of the standard into separate packages. However, this meant that users were obliged to consume the standard as a whole, in all its multilayered complexity and technical detail, even if they were only interested in specific parts. At the same time, several salient areas of linguistics/lexicographic description such as etymology were not covered at all in the original LMF plus the lack of modularisation/de-coupling made the (inevitable) addition of new parts awkward (especially given the ISO workflow for publishing materials). Moreover, the recommended XML-based serialisation for LMF turned out not to be sufficiently compatible with other leading markup standards, most prominently TEI-XML. For these reasons and others, the ISO sub-committee **ISO TC 37/SC 4/WG 4** was given the task of reviewing LMF with

[1] Electronic versions of print dictionaries or any electronic dictionaries which were originally intended for human consumption rather than for NLP tasks.

a view to creating a new version of the standard which would address these issues. The result is an updated version of the standard which, when all parts are published, will constitute a multi-part standard consisting of six separate modules, each published as a separate ISO standard, with further extensions planned to come. Most importantly, the new version of LMF will be backwards compatible with the 2008 version in order to ensure continuity and interoperability with lexicons encoded using the previous version.

In keeping with the fundamental conceptual modelling principles settled on by ISO TC 37/SC 4/WG 4, the new LMF 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). To summarise, we have carried out major improvements in the following areas: **restructuring**, **enrichment** and **simplification**. When it comes to **restructuring** the standard, we have already mentioned that the new version of LMF is being published as a multi-part standard in order to ensure a greater level of modularity. In terms of **enriching** LMF, we have introduced several new classes and properties amongst which **Bibliography** for specifying references for usages, definitions, examples, etc[2]. In general, 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[3].

- **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 (Khan and Bowers, 2020) 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.

References

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 per l'Informatica Umanistica e la Cultura Digitale*.

Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Bański. 2019. Lmf reloaded. *arXiv preprint arXiv:1906.02136*.

[2] One other important new novelty is the differentiation of Orthographic Representation into Form Representation and Text Representation has been designed to enable more precision in the encoding of written forms touching respectively Sense and Form sub-classes.

[3] A seventh part dealing with morphology is being planned as well as separate Metadata part.