Modelling Lexicographic Resources Using CIDOC-CRM, FRBRoo and Ontolex-Lemon

Anas Fahad Khan
Istituto di Linguistica Computazionale <<A. Zampolli>>, CNR, Italy
fahad.khan@ilc.cnr.it
Ana Salgado
NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal & Academia das Ciências de Lisboa, Portugal
anacastrosalgado@gmail.com

Goals

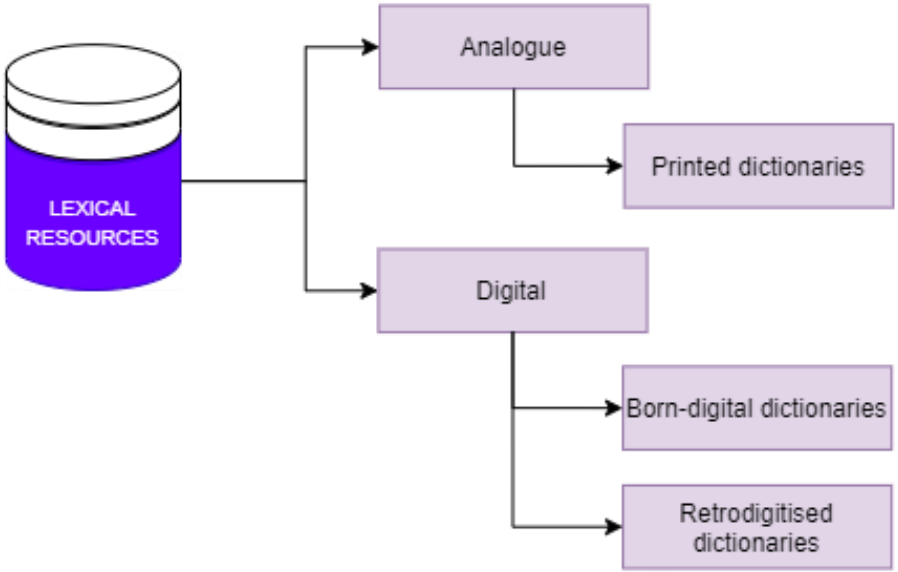
1. To present a **new approach** to the modelling and publication of lexicographic resources as linked data.
    a. Approach is based on the use of the **CIDOC-CRM** aligned **FRBRoo** ontology with **Ontolex-Lemon** model and its follow-up module, **lexicog**
    b. We propose new classes to act as **a bridge** between FRBRoo and Ontolex-Lemon
2. To present a case studies in the use of ontologies model texts as **complex/hybrid objects** at different levels of description.

Lexicographic Resources

- Focus on **lexicographic resources**: **digital editions** of **paper dictionaries** & lexical resources with **a dictionary-like interface**.
    - These are interesting for ***how*** they represent linguistic information and for the information **itself**.
    - Well known case of **retro-digitised dictionaries**: historic dictionaries converted to a **digital format**; there are interesting for the compilation process, revisions across editions, publishing histories, etc.
- Complex objects that are both **physical objects** and **informational content** (c.f., Pustejovksy's dot objects).
    - *The dictionary is outdated and very often incorrect in its etymological analyses but the definitions can be amusing and it makes a nice doorstop.*

Why Lexicographic Resources?

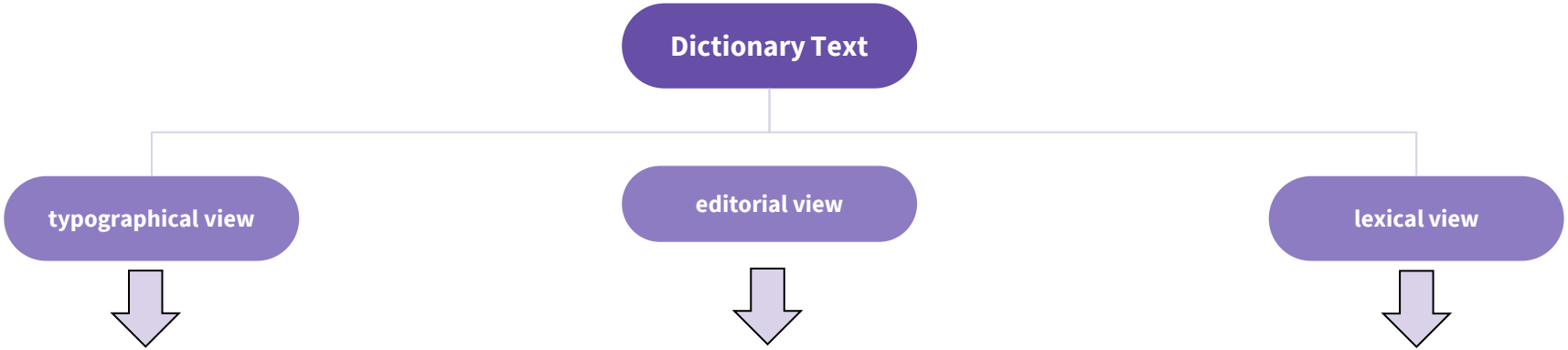
- The production of digital descriptions or digital versions of ***any*** kind of text confronts us with the distinction between the **content of a text**, the **content** is **presented** and **the history of the creation of the text**.
- Dictionaries are an interesting case: they organise similar kinds of linguistic information in **standardised ways**.
- Plus this linguistic content can be represented (in a formal way) much more easily than in other cases, e.g., plays, novels, encyclopedias, etc.
    - This makes them a useful test case in the modelling of texts using ontologies.


LEXICAL RESOURCES
Analogue
Printed dictionaries
Digital
Born-digital dictionaries
Retrodigitised dictionaries

Different Views on a Dictionary

- We look at how to model: the **visual appearance,** the **linguistic content** of lexicographic resources, along with other relevant historical and bibliographical facts in RDF using **Semantic Web vocabularies and models**.
- One of the best accounts of the different kinds of information to be taken into consideration when encoding dictionary texts is given by the **Text Encoding Initiative (TEI)** guidelines which provide for the formal modelling of text in **XML**.
- They are arranged in chapters/modules and allow for the markup of **structural and conceptual components of texts**. **Chapter 9** deals with dictionaries.

Different Views on a Dictionary



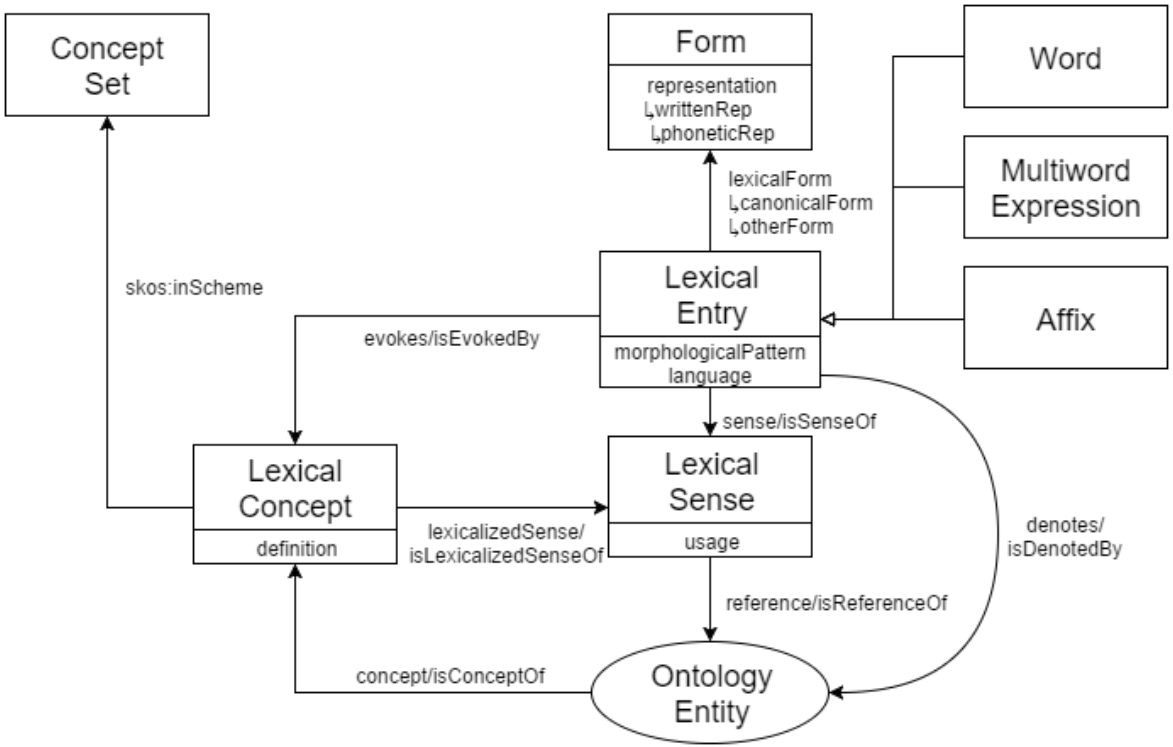
- “the two-dimensional printed page, including information about line and page breaks and other features of layout” (TEI P5)
- The properties of a text modelled as a sequence of tokens (words and punctuation), input to the typesetting process
- The conceptual or linguistic content of a dictionary as a whole, as well as its individual entries

Lexicographic Resources & the Semantic Web

- TEI already provides a means of representing different views on dictionaries and encoding these as XML.
- However…
  - If we want to publish lexicographic resources as **linked data** (and take advantage of the **Semantic Web stack** and especially formalisms such as **RDFS** and **OWL**), we will have to make everything much more explicit & machine actionable.
  - There is not a lot of **specific provision for lexicographic resources** vis a via Semantic Web ontologies and vocabularies.

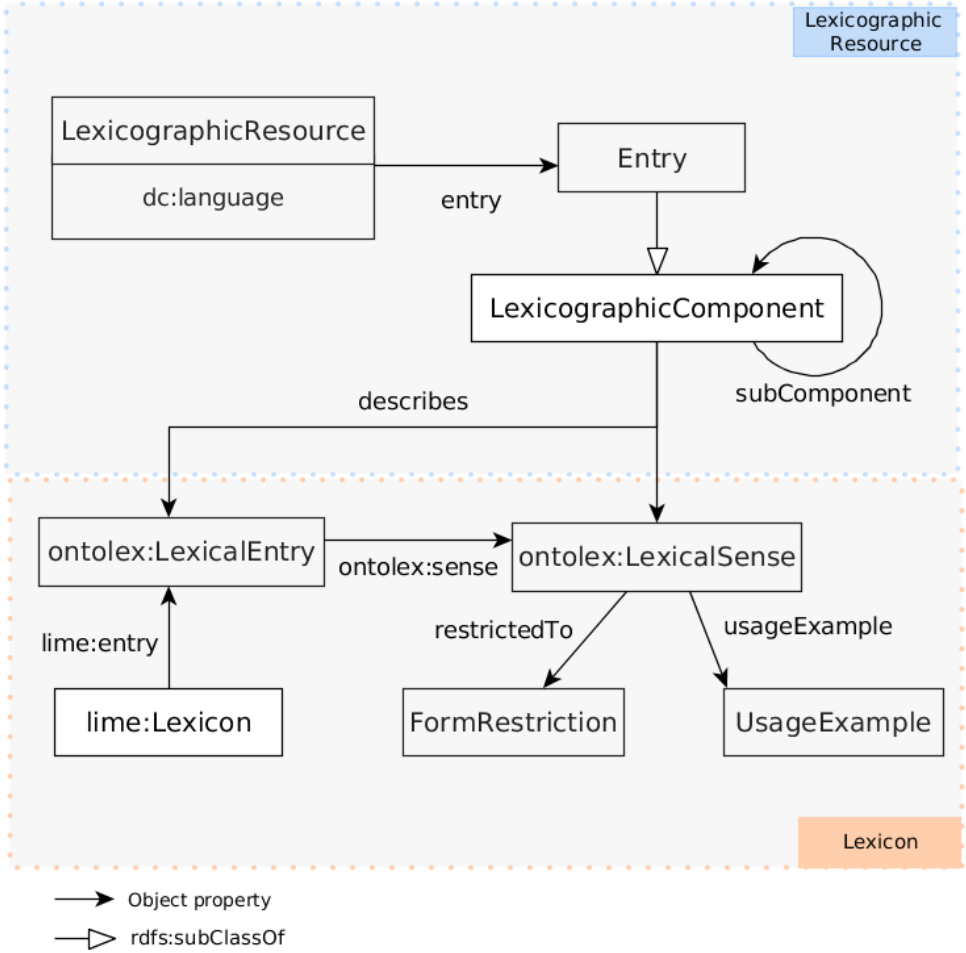
OntoLex-Lemon

- A popular Semantic-Web native model which allows for the modelling and the publication of **the lexical content** of lexicographic resources.
- Features such concepts as *Lexicon*, *Lexical Entry, Lexical Sense*, *Form*.
- Imposes a series of **restrictions** on how lexical content is represented which conflict with how many dictionaries represent their content.
- These restrictions mean that the lexical content which is represented is **rendered more interoperable**.


Concept Set
Form
representation
writtenRep
phoneticRep
Word
Multiword Expression
Affix
lexicalForm
canonicalForm
otherForm
skos:inScheme
Lexical Entry
morphologicalPattern
language
evokes/isEvokedBy
sense/isSenseOf
Lexical Concept
definition
Lexical Sense
usage
lexicalizedSense/
isLexicalizedSenseOf
denotes/
isDenotedBy
reference/isReferenceOf
concept/isConceptOf
Ontology Entity

Lexicog

- The **OntoLex-Lemon Lexicography Module (lexicog)** subsequently developed by the W3C OntoLex group to represent some of the structural information “lost” in an OntoLex-Lemon.
- It defines new classes such as **Lexicographic Resource** (complementing OntoLex **Lexicon**) which consists of single **Entry** individuals which represent lexicographic articles and which can be realised by OntoLex **Lexical Entry** elements.
- **Entry** is a subclass of **Lexicographic Component** which represents elements which describe the structuring of lexicographic articles.


Lexicographic Resource
LexicographicResource
dc:language
entry
Entry
LexicographicComponent
subComponent
describes
ontolex:LexicalEntry
ontolex:sense
ontolex:LexicalSense
lime:entry
restrictedTo
usageExample
lime:Lexicon
FormRestriction
UsageExample
Lexicon
Object property
rdfs:subClassOf

Dictionaries as Textual/Material Objects

- OntoLex-Lemon + Lexicog however **still aren't** sufficient to represent all the different aspects we might be potentially interested in.
  - Who **compiled** the dictionary, is it based on **previous works**?
  - What about the **publishing history** of the text itself, its **different editions** (with different entries, definitions, etc), its **translations, manuscripts,** what about **individual copies in libraries**?
  - What about the **texts/corpora** that are **cited as attestations**, citations to scholarly works?
  - For some of these there already exist generic vocabularies (**Dublin Core**, **Prov-O**, **CITO**) which can provide solutions, others have to be adapted to the dictionary domain.
- **FRBR** will provide us with a conceptual framework for integrating together different levels of description.

FRBR

- Stands for **Functional Requirements for Bibliographic Records**: an entity relationship model intended for the classification of intellectual products in **bibliographic databases** and **library catalogues**.
- It introduced an important distinction in terms of how we can describe intellectual products. We can refer to such products at four different levels of description. Namely, at the level of **Work**, **Expression**, **Manifestation**, and **Item**.
- We use the version of this distinction given in the **CIDOC-CRM aligned FRBRoo** ontology.

Work and Expression

- **F1 Work**: "*[C]omprises distinct concepts or combinations of concepts identified in artistic and intellectual expressions [...] The substance of Work is ideas*".
  - Note that in the case of dictionaries this would encompass the **TEI lexical view**.

- **F2 Expression**: "*[C]omprises the intellectual or artistic realisations of works in the form of identifiable immaterial objects, such as texts, poems [...] or any combination of such forms that have objectively recognisable structures*".
  - In the case of dictionaries we claim that this description encompasses the **TEI editorial view**.

Manifestation and Item

- Originally one class in the FRBR model, **Manifestation**, this latter corresponds to two separate classes in FRBRoo: **F3 Manifestation Product Type** and **F4 Manifestation Singleton**. The former class is said to define *" all of the features or traits that instances of F5 Item normally display in order that they may be recognised as copies of a particular publication"* ; the latter as *"[comprising] physical objects that each carry an instance of F2 Expression, and that were produced as unique objects"*
  - In the case of dictionaries F3 Manifestation Product Type encompasses the **TEI typographic view**.
- The **Item** class : "*[C]omprises physical objects*" such as specific physical copies of dictionaries kept at libraries or academic institutions.
  - This class is associated with the kind of metadata information that is usually contained within the **TEI header element**.

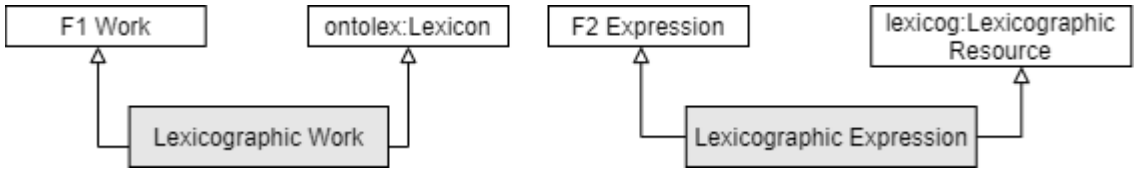
FRBRoo and Lexicographic Resources

- Take a **multi-edition dictionary**, we can represent it as a **F15 Complex Work**.
- Different individual editions classed as instances of **F15 Individual Work**.
- Each of these different editions can then be described at the level of **F2 Expression** in order to specify, for example, the wording of individual entries.
- Moreover we can also describe the dictionary at the level of **F3 Manifestation Product Type** in order to specify the content and placement of images and their relation to the text (this is important in the case of **illustrated dictionaries**).
- **FRBRoo** also allows for the modelling of dictionaries which have been translated from one language to another.

Bridging FRBRoo and OntoLex

- We propose a number of new classes and properties to bridge together FRBRoo (and CIDOC-CRM) and OntoLex-Lemon.
- **Lexicographic Work:** A subclass of the FRBRoo class **F1 Work** and the Ontolex-Lemon class **Lexicon**. It comprises concepts or combinations of concepts for representing/describing the lexicon for a given language community or communities or domain.
    - As **F1 Work** is a subclass of the CIDOC-CRM class **E89 Propositional Object** we can view individuals of **Lexicographic Work** as sets of **propositions *about* lexemes and related linguistic concepts belonging to a lexicon**.
- **Lexicographic Expression:** A subclass of the FRBRoo class **F2 Expression** and the lexicog class **Lexicographic Resource:** The class comprises an intellectual realisation of the description of a lexicon as a structured text.
    - In other words it is a text viewed apart from **a specific typographic realisation**: a sequence of words that has an **additional organisation** in terms of entries, senses (defined as a sub-part of a lexicographical article that discusses a meaning of a lexical unit), forms, etc.

Bridging FRBRoo and OntoLex



Asserting the Lexical View

- In our approach, we view a lexicographic article as a series of statements making claims about different linguistic phenomena, about the lexicon of a language, as well a structural component of a text. In this we elaborate on previous work in both OntoLex and in CIDOC/FRBRoo.
- By modelling a dictionary as consisting of different levels of information, we can explicitly represent these as **hypotheses** (using named graphs or nanopublications).
- This comes in especially useful when it comes to combining together **etymologies**.

Summary and Conclusions

- The use of ontologies helps to improve the interoperability and re-usability of the lexicographic resources which are in reality complex hybrid objects. It makes these aspects of the resource much more accessible.
- TEI doesn't offer us the same amount of expressivity (though it has other advantages).
- So far we have studied different types of lexicographic use cases (including modelling differences across editions, combining etymological information).
- The authors of the current work are involved in the digitisation of a historic dictionary *Diccionario da Lingua Portugueza* by Antonio de Morais Silva, as part of a Portuguese national project, *MORDigital*.
- We plan to apply the ideas presented in this work to the RDF version of this the digital edition of this dictionary.