One Model, Many Languages?

An approach to developing global language content

Euralex, Ljubljana, July 2018

Judy Pearsall
Oxford University Press
Oxford Global Languages

Key areas:

1. Technology
   • Central platform
   • Internal and external use cases
2. Methodological
   • Single lexicographical framework
   • Appropriate tools
3. Ways of working
   • Iterative, Agile approach
   • Strong collaboration editorial/technology
4. Data representation
   • Single data model
   • Syntactic/semantic interoperability
The vision:
To bring lexical content online for 100 of the world’s languages and make it available to developers, consumers, licensees, and researchers for a wide variety of uses

The mission:
To improve the quality and breadth of global linguistic knowledge and communication, giving voice to all people in a rapidly changing world
Oxford Global Languages

Dr Langa Khumalo at the launch of Oxford Global Languages on 2 September 2015
Oxford Global Languages
So far...

18 languages launched

- Gujarati
- Hindi
- Indonesian
- isiXhosa
- isiZulu
- Latvian
- Malay
- Northern Sotho
- Romanian
- Setswana
- Southern Quechua
- Swahili
- Tajik
- Tamil
- Tatar
- Tok Pisin
- Turkmen
- Urdu

Oxford Global Languages

https://developer.oxforddictionaries.com
Scaling Oxford Global Languages

- 30 languages by March 2020
- Focus on Indian languages
- New architecture and Data Model
- Content creation model
- Experimental and Agile
What do we mean by one model?

A standardized representation of lexical data and language data

Allows terminology, tools, and outputs to be interoperable between languages

Scalable to new languages and data features
Why one model?

- Answers needs of our customers
- Single system and set of tools
- Cost efficiency
- Multilingual linking
- Efficient data management
Main challenges and objections

- Size of task
- Complex legacy data from disparate sources
- Languages are not the same!
  - Alignment can flatten valid distinctions
  - Potential for false analogy
- Dealing with different lexical traditions
- Tendency for English/European languages to predominate
Dictionaries Domain Model

How our data is used: use cases

Domain Model

Rules for data features

Data model

Original data
The print business

Each dictionary is a data silo
Our business has changed

- Data
- Websites
- Mobile apps
- Virtual Assistants
- AI
- Social media
- IOT
- Chatbots
- Email
- Print
The market shift
Language Data and Services Sector (not to scale)

DICTIONARIES IN THE TRADITIONAL SENSE

RELATED LEXICAL DATA

CORPORA, TRAINING DATA

THE LANGUAGE BUSINESS (~£4.5bn annual)
Domain model

A map of the domain to help solve real-world business, data, software, and design problems.
Domain model

Lexical Entry

Lexical Form

Display Form

Homograph ordering

Normalized Form

Lexical Item

A conceptually distinct word or multi-word expression
- e.g., bear (animals) or bear (to carry)
- bear in mind, bear up, bear hug

Transliterations

May also occur at morphemic and Sense Level
Different data sources
Domain Model: single set of terminology

- Lemma
- Headword
- Label
- Part of Speech
- Entry
- Word

Lemma

Lemma
Domain Model: print structures

Space-saving treatments in print

Parts of speech combined

Words embedded within another item

elate

VERB

[WITH OBJECT]

(usually as adjective elated)
Make (someone) ecstatically happy.

beta rhythm

NOUN

[MASS NOUN] Physiology

The normal electrical activity of the brain when conscious and alert, consisting of oscillations (beta waves) with a frequency of 18 to 25 hertz.
Domain Model: print structures

Space-saving treatments in print

Cross references to other words
Domain Model: derivatives and compound forms

Asylum

Definition of asylum in English:

**asylum**

NOUN

1. (also political asylum) [mass noun] The protection granted by a state to someone who has left their home country as a political refugee.

Political asylum

*political asylum* n. asylum granted by a country to a political refugee from another country.

1852 *Times* 28 Feb. 4/4 These cantons, with their free press, their political asylum, and their creed, are intolerable to the jealous eye of neighbouring despotism.

Asilo

Translation of asilo in English:

masculine noun

(proteção)

asylum

Asilo político

asilo político

Political asylum

Translation of political asylum in Russian:

политическое убежище

political asylum
Lexical entry, Lemma

sausage

m_en_gbus 0902490

(NOUN)
Lexical entry, Lemma + Morphology

sausage

sausage_ NN_1

sausage_ NN_1.1

(NOUN, singular)

sausage_ NN_1.2

(NOUN, plural)

m_en_gbus 0902490

(NOUN)

sausages

Lexical Entry
MorphUnit
Paradigm
Inflection
Wordform
MSCat
Has Lexical Entry
Lemma/Display
Form
Normalized
Has MsCat
Has Inflection
Has Wordform
Lexical entry, lemma + Variant

colour

m_en_gbus 0200980
(NOUN)

m_en_gbus 0200980-se1-2
(VERB)

color

MSCat

inflection

wordform

MSCat

has lexical entry

has lemma/displayed form

has normalized form

has ms cat

has inflection

has wordform
Domain Model Terminology

• **Lexical Item**
  – The abstract concept of a “word” or “discrete linguistic item” as distinct from the canonical spelling of it

• **Display Form**
  – the *wordform* typically used to write a **Lexical Item**

• **Lexical Form**
  – The combination of wordform and homograph number that uniquely identifies a **Lexical Item** (e.g. "march 1")
### Tonality

**Igbo:**
a standardized code is used

**Yoruba:**
a system of accents is used as part of the translation
Domain Model: corpora

**Wordform**

**Meaning potential**

- Potential new word candidates

**Wordform patterns**

**Adding context**

- Absolute frequency
- Relative frequency
- Regional distribution
- Grammatical formation
- Electronic trends
- n-grams
- Collocations

**Morphology**

- Display Form
- Inflections (e.g., coloured, colouring)
- Paradigms (e.g., Inflections of colour, n.)
- Lemma
Basic Rules for All Dictionaries

9. DISPLAY FORMS - BRAD

GENERAL PRINCIPLES

1. Schemas should change rarely
2. Controlled vocabularies should be used and adhered to
3. The same tag names should be used for every dataset
4. Mixed content should be avoided
5. Every tag / object / item should be one thing and only one thing

What is a display form?

A formal representation of the lexical item in a particular language, used as the header of a particular lexical entry. This may be the way in which it is typically written in a language, or it may differ (for example Arabic display forms will have vowels).

Each lexical entry has its own display form and has only one display form. For example, "march" has the display form "march", and "march" has the display form "march".

A lexical entry can only have one display form, but there may be variant forms (e.g. color as a variant of colour), which we might choose to display instead of the display form for particular use cases. We will discuss this further under "variants".

12. HOMOGRAPHS - BRAD

GENERAL PRINCIPLES

1. Schemas should change rarely
2. Controlled vocabularies should be used and adhered to
3. The same tag names should be used for every dataset
4. Mixed content should be avoided
5. Every tag / object / item should be one thing and only one thing

What is a Homograph?

- A homograph is lexical item that shares the display form with another lexical item, e.g. bank "money institution" versus bank "river bank".

- A near homograph is a lexical item that shares a normalised form with another lexical item, but differs in upper/lower case, diacritics, punctuation. E.g. polish vs. Polish, rose vs. rosé, Urdu جہاں جہاہ‎ vs. جہاں جہاہ. am vs. a.m., resign vs. re-sign
  - In abugidas (e.g. most Indian scripts), display forms with differences in diacritics are treated as entirely separate lexical items, not as homographs or near-homographs. E.g. Hindi जब and जेब.
Ex. 1) If a variable display form is provided, the data must allow expanded versions of all implied lexical forms.

**swipe right (or left)**

**PHRASE**

**informal**

(on the online dating app Tinder) indicate that one finds someone attractive (or unattractive) by moving one’s finger to the right (or left) across an image of them on a touchscreen.

‘I swiped right, but sadly for me, she swiped left’

**figurative** ‘are elephants more likely to ‘swipe right’ when they see mates with longer trunks?’

Ex. 2) Separate words with multiple morphosyntactic categories.

**al fresco**

**ADVERB & ADJECTIVE**

In the open air.

[as adjective] ‘an al fresco supper’

[as adverb] ‘in the unlikely event of some sunshine you can even dine al fresco’
Master Data Model

What is it?

• A data model for JSON, applicable to all types of lexical data.

• Takes the Domain Model as its base, and follows the principles of the defined Rules.

• Covers monolinguals, bilinguals, thesauruses, morphology, wordlists
Data harmonization, part 1

Print level outputs

Data wrangling

Syntactic –level conversion

Original data
Data harmonization, part 2

Multiple level outputs

- Semantic-level conversion
- Syntactic-level conversion

Original data

Domain Model, Use Cases and Rules
New editing tools
Thank you

With special thanks to Khalil Ahmed, Imogen Foxell, Marie-Claire Holmes, Ruth Madder, Roser Sauri, Richard Shapiro, Angus Stevenson and all the Oxford Global Languages team

Judy Pearsall
Oxford University Press
Appendix