



KNOWDIVE



KDI ● **Knowledge and Data Integration**

Language Alignment

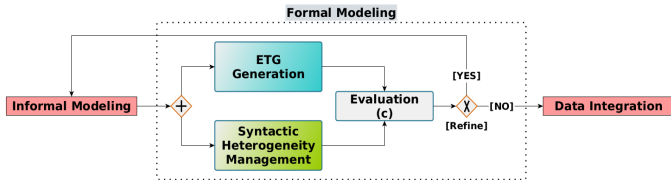
iTelos Formal Modeling Phase

Fausto Giunchiglia, Mayukh Bagchi

Contents

- 1** *iTelos* Formal Modelling Phase - Overview
- 2 Language Alignment - Theory
- 3 Language Alignment - Process
- 4 Summary

iTelos Phases - Formal Modeling



Inputs:

- ETG Model.
- Selected Datasets.
- Reference Ontologies

Outputs:

- ETG.
- Datasets syntactically formatted.

Objectives:

- **Knowledge layer:** Create a shareable ETG, following the ETG model and reusing semantically synonymous concepts from reference ontologies
- **Data layer:** The datasets selected are, in this phase, elaborated through the DTA-2.1 (See iTelos Principles - slide 6).

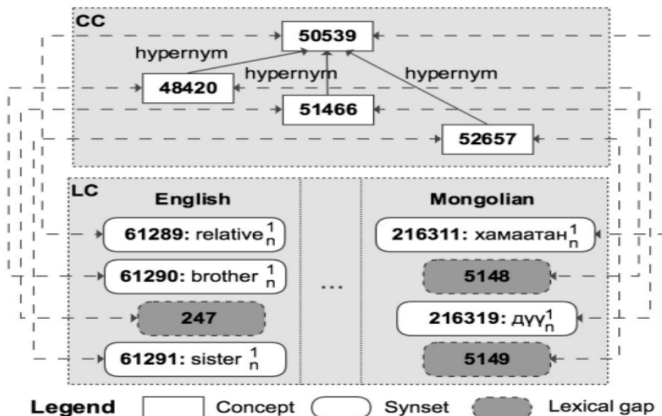
Contents

- 1 *iTelos* Formal Modelling Phase - Overview
- 2 Language Alignment - Theory**
- 3 Language Alignment - Process
- 4 Summary

What We Have

- An *informal* ETG Model. *Informal* because:
 - 1 The concepts in the ETG Model are still not *uniquely* and *formally* expressed, thus exhibiting conceptual diversity (L1)
 - 2 The concepts in the ETG Model are still expressed using *natural language words*, thus exhibiting language diversity (L2)
 - 3 Due to the persistence of conceptual diversity (point [1] above) and language diversity (point [2] above), it is still not possible to adapt the ETG Model, for instance, to a different language or culture.

UKC Architecture



Giunchiglia, Batsuren, Freihat. In Proceedings of CiCling (2018)

Definitions

- 1 **Synset:** Multiple meanings of a word are codified as a lexicalized concept, also called a synset, consisting of a (possibly incomplete) set of synonymous words.
- 2 **Word Sense:** In linguistics, a word sense is one of the meanings of a word. For example, the word *play* has 50 different meanings.
- 3 **Concept Core (CC):** The Concept Core (CC) is the UKC representation of the world and it consists of a semantic network where the nodes are language independent concepts.

Definitions

- 1 Language Core (CC):** We talk of the Language Core (LC), meaning the component that, in the UKC, corresponds to the set of words, senses, synsets, glosses and examples supported by the UKC.
- 2 GID:** Global Identifier (GID) is the unique identifier assigned by the UKC CC to a concept.
- 3 Princeton WordNet (PWN):** [PWN](#) is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

Language Alignment

Language Alignment transforms the *informal concepts* of the ETG Model to *formal concepts* with the *key* support of the UKC. Three main objectives:

- 1 Representing each informal concept in the ETG Model with the unique identifier (GID) from UKC, thus, rendering each concept *formal* and absorbing L1 diversity.
- 2 The ETG Model vocabulary has both *already existing concepts* in the UKC CC, and, *new concepts* absent in the UKC CC. We align the existing concepts with their equivalents in the UKC CC and extend the UKC CC by adding (only) the new concepts (with new GIDs), thus, absorbing L1 as well as L2 diversity.
- 3 Due to the LC, each concept can also be rendered multilingually, thus opening up the possibility to *adapt* the ETG in any language or culture.

Contents

- 1 *iTelos* Formal Modelling Phase - Overview
- 2 Language Alignment - Theory
- 3 Language Alignment - Process**
- 4 Summary

Process Input and Output

- Process Input:

- The (informal) Entity Type Graph (ETG) Model
- UKC Knowledge Organization System [UKC KOS] Application (the instance for your project)

- Process Output:

- Each concept in the ETG Model identified formally by a unique GID
- UKC CC aligned with the existing concepts, and extended with the new concepts from the ETG Model

Methodology

The general methodology for language alignment, as semi-automatically performed via KOS application, is as follows -

- Each term (concept) is selected from amongst the classes, relations and attributes (all terms from all these hierarchies are considered; one at a time)
- The term is (semantically) searched in the UKC Knowledge Base (UKC KB) via the KOS (iTelos) application, and the step will result in *one* of the following *two* scenarios:-
 - 1 Scenario 1 (*S1*): *Synonymous Match* between the ETG concept and a synonymous concept found in the UKC KB
 - 2 Scenario 2 (*S2*): *No Existing Semantic Match* between the ETG concept and any concept in the UKC KB (that means, New Concept !!!)

Methodology (Contd.)

(S1): In case of a *Synonymous Match*, the concept is mapped to the semantically synonymous concept in the UKC CC, and thus is assigned the latter's GID. The key step is to choose the semantically synonymous concept amongst the options suggested by the KOS application.

(S2): In case of a *No Existing Semantic Match*, the concept, via the KOS application, is declared as a new concept with respect to the UKC CC.

- This step is *incremental* as all concepts from all hierarchies are sequentially examined.
- This step generates a spreadsheet file where each new concept of the ETG Model is recorded as a negative integer in a decremental negative integer sequence.
- In addition, the parent concept (and its GID) of the considered concept as in the ETG Model hierarchy is also recorded

Methodology (Contd.)

Finally, the spreadsheet file as generated above is automatically imported into the UKC Knowledge Base. This step results in:

- 1 alignment of all the synonymous matching concepts with their equivalent counterparts in the UKC KB, and thus each concept annotated with a unique GID
- 2 extension of the UKC Knowledge Base with the new concepts. The '-ve' IDs are automatically replaced with a new GID

The process of Language Alignment sets the basis for the schema alignment, where we map the ETG Model to the Foundational Teleology, generating the final, fully formal ETG.

Spreadsheet Example - schema.org

	A	B	C	D	E		
1	Case	Word Lemma	Description	Concept UK ID	PoS	Parent Concept + Parent UK ID	Relation Kind
2	schema	DataType	The basic data types such as Integers	-1	n	Class 43482	IS_A
3	schema	Boolean		97088	n	schema:DataType -1	IS_A
4	schema	Date		80741	n	schema:DataType -1	IS_A
5	schema	DateTime	A combination of date and time of day	-2	n	schema:DataType -1	IS_A
6	schema	Number		36475	n	schema:DataType -1	IS_A
7	schema	Float		103163	n	schema:Number 36475	IS_A
8	schema	Integer		72941	n	schema:Number 36475	IS_A
9	schema	Text		34287	n	schema:DataType -1	IS_A
10	schema	CssSelectorType	Text representing a CSS selector.	-3	n	schema:Text 34287	IS_A
11	schema	URL		34123	n	schema:Text 34287	IS_A
12	schema	XPathType	Text representing an XPath (typically by	-4	n	schema:Text 34287	IS_A
13	schema	Time		80563	n	schema:DataType -1	IS_A
14	schema	Thing		25691	n	schema:Class 43482	IS_A
15	schema	Action		161	v	schema:Thing 25691	IS_A
16	schema	AchieveAction		152	v	schema:Action 161	IS_A
17	schema	LoseAction		105645	v	schema:AchieveAction 152	IS_A
18	schema	TieAction		74141	v	schema:AchieveAction 152	IS_A
19	schema	WinAction		39532	v	schema:AchieveAction 152	IS_A
20	schema	AssessAction		103695	v	schema:Action 161	IS_A
21	schema	ChooseAction		103662	v	schema:AssessAction 103695	IS_A
22	schema	VoteAction		112526	v	schema:ChooseAction 103662	IS_A
23	schema	IgnoreAction		105439	v	schema:AssessAction 103695	IS_A
24	schema	ReactAction		103860	v	schema:AssessAction 103695	IS_A
25	schema	AgreeAction		105311	v	schema:ReactAction 103860	IS_A
26	schema	DisagreeAction		104261	v	schema:ReactAction 103860	IS_A
27	schema	DislikeAction		40352	v	schema:ReactAction 103860	IS_A
28	schema	EndorseAction		112460	v	schema:ReactAction 103860	IS_A
29	schema	LikeAction		109148	v	schema:ReactAction 103860	IS_A
30	schema	WantAction		40264	v	schema:ReactAction 103860	IS_A

Contents

- 1 *iTelos* Formal Modelling Phase - Overview
- 2 Language Alignment - Theory
- 3 Language Alignment - Process
- 4 **Summary**

Summary

- We learnt about the motivation behind the language alignment activity of the *iTelos* methodology
- We understood the underlying methodology for aligning and extending the UKC knowledge base
- We saw an example of the spreadsheet file for the language alignment process performed on schema.org



KDI : Knowledge and Data Integration



Fausto Giunchiglia, Mayukh Bagchi



Language Alignment
iTelos Formal Modeling Phase