Natural Language Processing

The Arabic Ontology

Mustafa Jarrar
Birzeit University
Palestine

Jarrar©2025

Watch this lecture
and download the slides

Course Page: http://www.jarrar.info/courses/NLP/
More Online Courses at: http://www.jarrar.info

Keywords: Arabic, Arabic NLP, NLP, Natural Language Processing, Arabic Ontology, Linguistic Ontologies, Wordnets, Machine Translation, Word Sense Disambiguation, Semantic Web, Knowledge Graphs, Graph Neural Networks, Ontologies, Application Ontology, Tree of Meanings, Formal Arabic WordNet, Formalisms, Morphology Level, Semantic Level, Synset, Concepts, Semantic Relations, Wikidata, Upper-level Concepts, Disjointness, Gloss, Synonyms, Hyponymy, Implied Relation, Artificial Intelligence, Deep Learning, Corpus, Corpus Linguistics, Multilingualism, Lexical Semantics, EURO WordNet, Wordnet, Arabic WordNet, Global WordNet, Antonymy, Hyponymy, Meronymy, Thesauri, DOLCE, BFO, Upper-Level Ontology, Qabas, Lemma, Example, Download, Entity, Object, Occurrent, Dependent Entity, Abstract, Formal Entity, Informatio, الأنطولوجيا، حوسبة اللغة، مكنز، تعريف، مفهوم، معجم ذهني، مترادفات، وردنت، حوسبة الدلالة، شبكة المفردات، العربية، أنطولوجيا لغوية، تعدد اللغات، الترادف اللغوي، تعدد المعاني، التضاد، تصنيف المعاني، علاقات جزء-كل، شجرة المعاني، شجرة الأنطولوجيا، شبكة معرفة، مفهوم، كليات، ماصدق، أنطولوجيا، إبستمولوجيا، ترادف، مترادفات، حدود، حقل دلالي، شجرة مفاهيم، تصنيف، صفات جوهرية، ويكي بيانات، قبس، مفتوحة المصدر، مثال، تنزيل، كينونة، موجود، سيرورة، منوط، مجرد، منوط.

Natural Language Processing

The Arabic Ontology

In this lecture:

- Part 1: **Need for Linguistic Ontologies**
- Part 2: General Overview
- Part 3: Upper-level Concepts
- Part 4: Gloss Formulation Guidelines
- Part 5: Fundamentals and Formal Definitions
- Part 6: Benchmarking Ontology Content
- Part 7: Discussion
- Part 8: Practice

Reading

Everything in these slides

Mustafa Jarrar: **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content**. Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021

https://www.jarrar.info/publications/J21.pdf

Mustafa Jarrar: **Linguistic Ontologies and Wordnets**. *Invited Speaker*. The 11th Global Wordnet Conference. South Africa 18/1/2021

https://www.youtube.com/watch?v=Pgf4MzTHJc4&list=PLSu_tNkEO5Y1fW4VLR1GXv6yvSHZaHLLm&index=21

➔ See the references at the end.

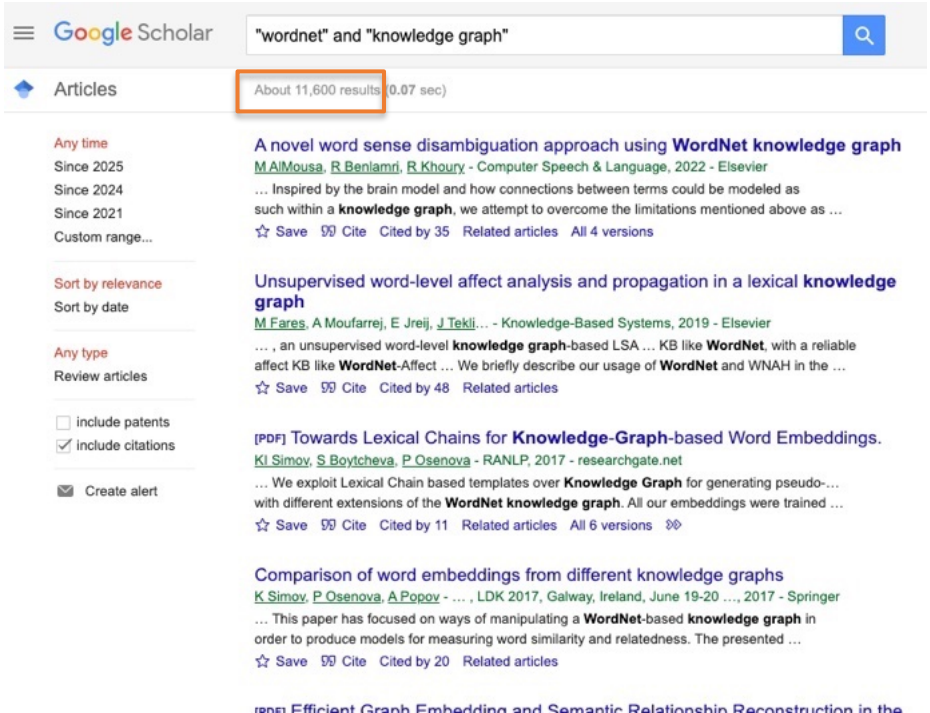
Why Linguistic Ontologies

❖ Typically, wordnets and linguistic ontologies are used in understanding and retrieving unstructured information in many NLP in IR tasks, such as:

- **Information Search and Retrieval** - to enrich queries and improve the quality of the results, i.e., meaningful search rather than string-matching search; cross lingual IR, etc.
- **Machine Translation, Word Sense Disambiguation …** – as a sense inventory to determine the intended meaning of a term within a context;
- **Classification** –to classify information and documents.
- **Data Integration and Interoperability** - as a semantic reference to several autonomous information systems;
- **Semantic Web** - as a semantic reference to disambiguate the meanings used in the web sites;
- among many, **many other applications.**

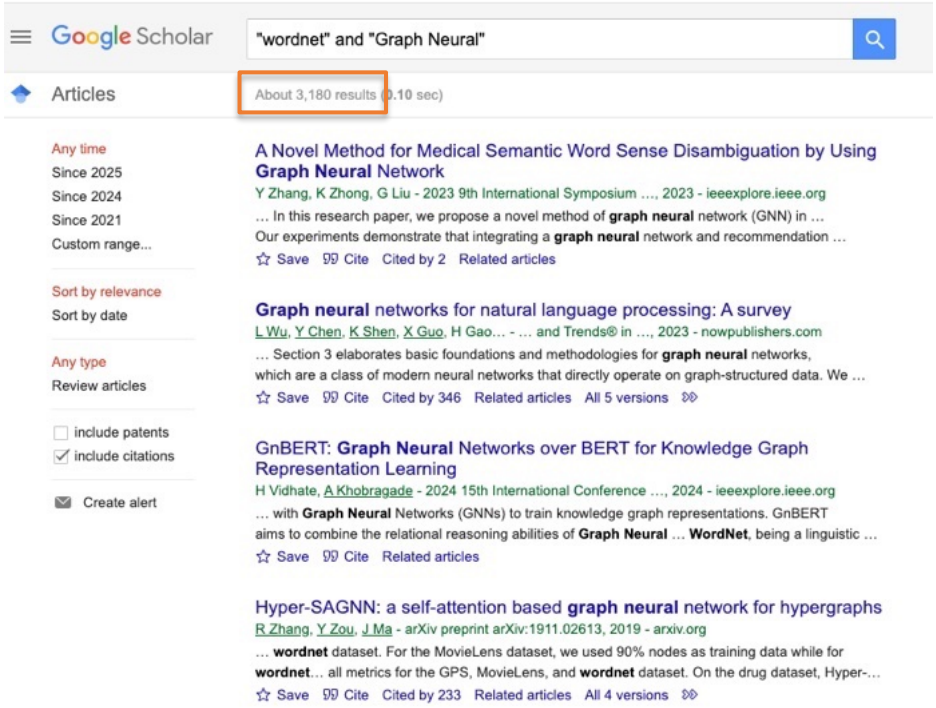
Why Linguistic Ontologies

- Typically, wordnets are used in understanding and retrieving unstructured information in many NLP in IR tasks, such as smart information retrieval, word sense disambiguation, classification, translation, and data integration.
- New demands are emerging to use wordnets like ontologies: to manage and retrieve structured data in e.g., cross lingual Big Data, and medical informatics, Knowledge Graphs, Graph Neural Networks, etc.



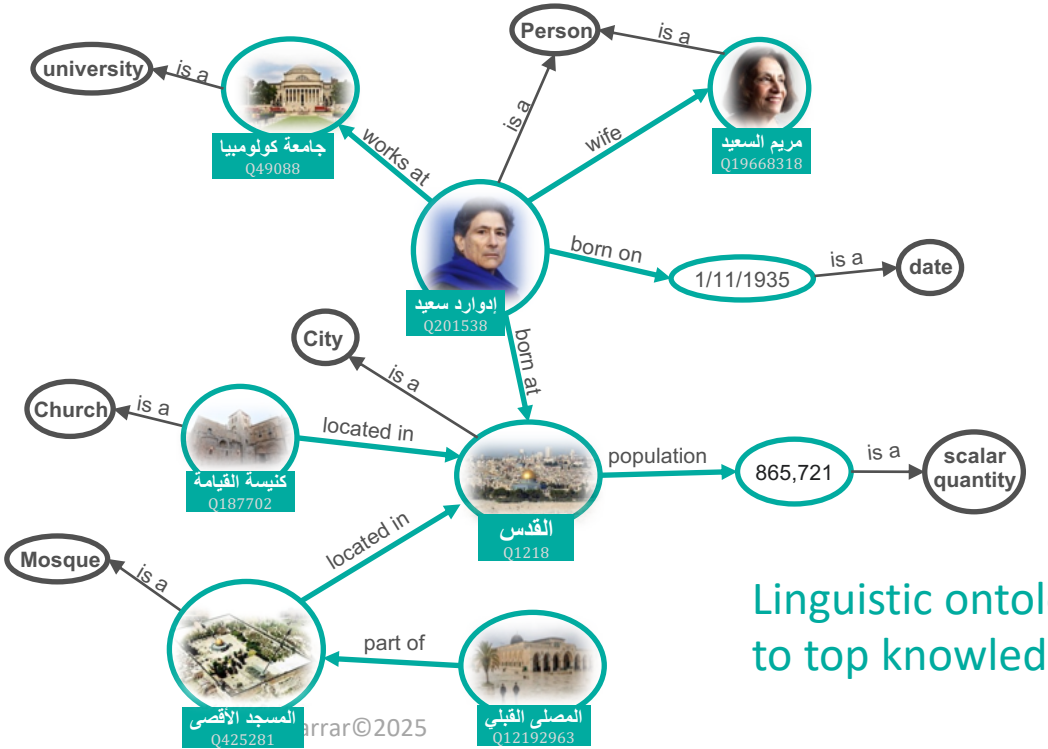
Why Linguistic Ontologies

- Typically, wordnets are used in understanding and retrieving unstructured information in many NLP in IR tasks, such as smart information retrieval, word sense disambiguation, classification, translation, and data integration.
- New demands are emerging to use wordnets like ontologies: to manage and retrieve structured data in e.g., cross lingual Big Data, and medical informatics, Knowledge Graphs, Graph Neural Networks, etc.



Why Linguistic Ontologies

- Typically, wordnets are used in understanding and retrieving unstructured information in many NLP in IR tasks, such as smart information retrieval, word sense disambiguation, classification, translation, and data integration.
- New demands are emerging to use wordnets like ontologies: to manage and retrieve structured data in e.g., cross lingual Big Data, and medical informatics, Knowledge Graphs, Graph Neural Networks, etc.

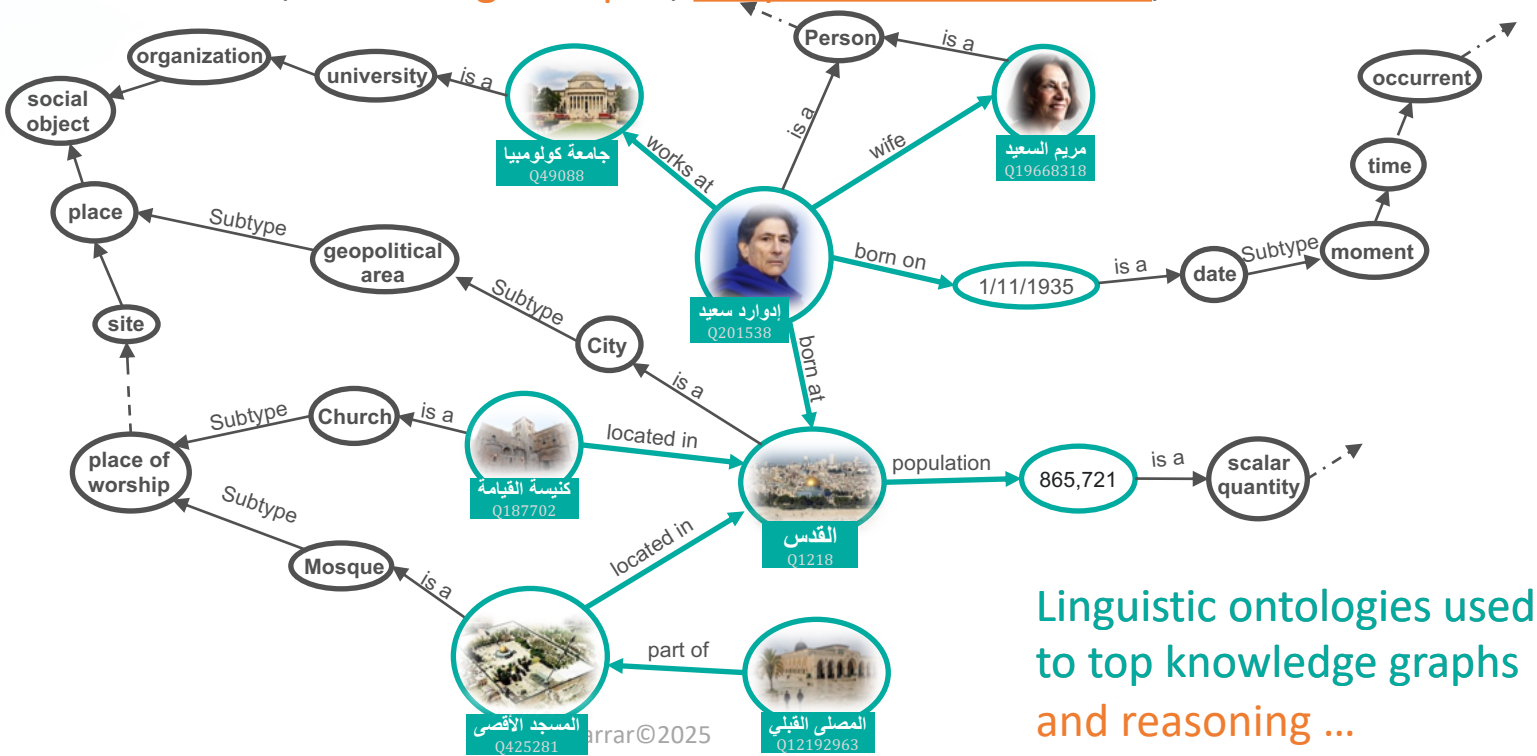


Linguistic ontologies used to top knowledge graphs

arrar©2025

Why Linguistic Ontologies

- Typically, wordnets are used in understanding and retrieving unstructured information in many NLP in IR tasks, such as smart information retrieval, word sense disambiguation, classification, translation, and data integration.
- New demands are emerging to use wordnets like ontologies: to manage and retrieve structured data in e.g., cross lingual Big Data, and medical informatics, Knowledge Graphs, Graph Neural Networks, etc.



Linguistic ontologies used to top knowledge graphs and reasoning …

arrar©2025

Why Linguistic Ontologies

But,
Ontologies are typically application-specific rich axiomatizations;
Wordnets are general-purpose mental lexicons, and thus axiomatizing wordnet would be a **rigidification**.

- How to build a linguistic ontology as a wordnet - to better serve new application scenarios
- The Arabic Ontology is a linguistic ontology (or a formal Arabic WordNet)

Application Ontology vs Linguistic Ontology

- Typically rich axiomatization
- Each term refers to one concept (no polysemy).
- Synonymy is not a target.
- Benchmarked to application's knowledge.
- Used by a certain application or a class of applications.

- Light-weight axiomatization, and cannot be rigid.
- Each term refers to one or more concepts (Polysemy).
- Synonymy is important.
- Benchmarked general knowledge
- Used for general purposes.

Jarrar©2025

Natural Language Processing

The Arabic Ontology

In this lecture:

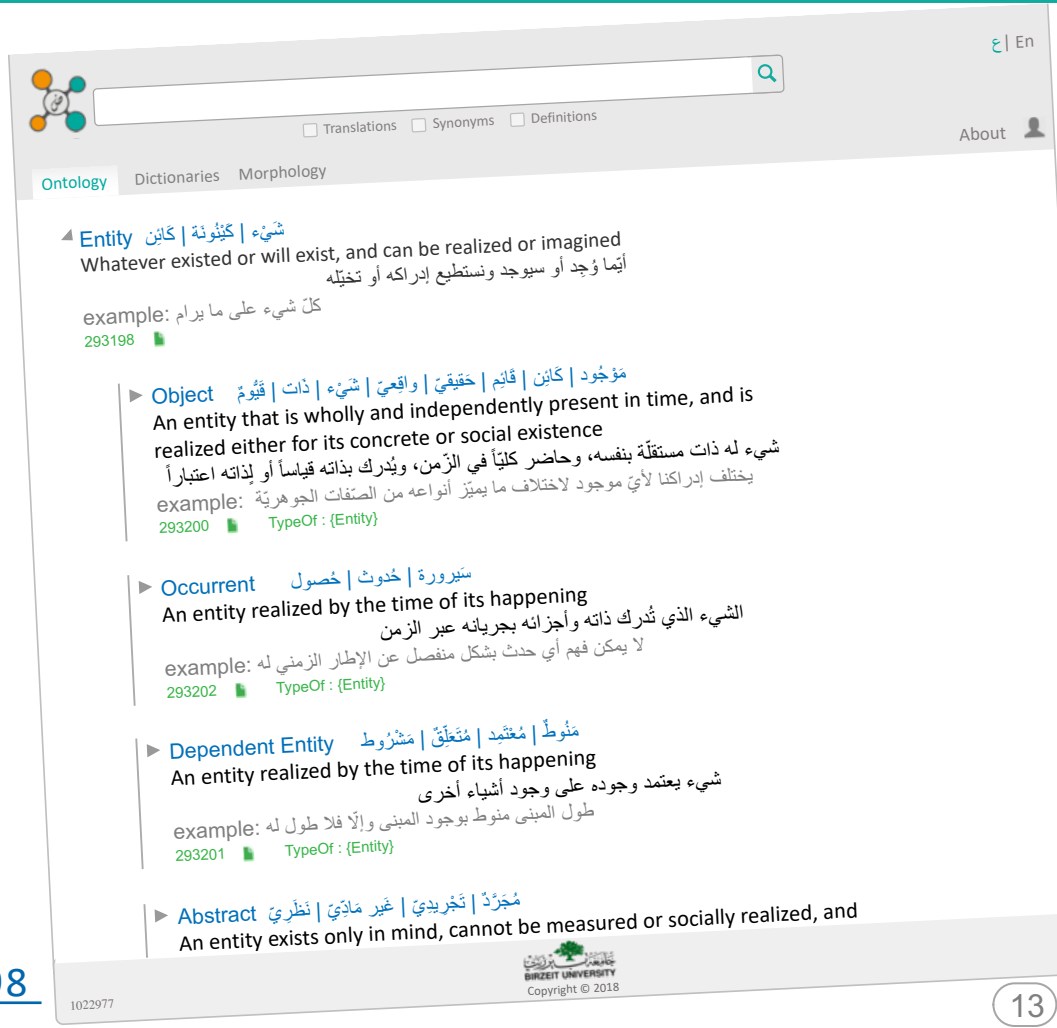
- Part 1: Need for Linguistic Ontologies
- Part 2: **General Overview**
- Part 3: Upper-level Concepts
- Part 4: Gloss Formulation Guidelines
- Part 5: Fundamentals and Formal Definitions
- Part 6: Benchmarking Ontology Content
- Part 7: Discussion
- Part 8: Practice

Arabic Ontology

- Characterization of the intended meaning (i.e., concepts) that the Arabic words convey.
- **Tree of meanings:** classifying concepts is a good way to specify these meaning.
- The classification is **based on intrinsic and distinguishing properties** (using the formal subsumption relation).

توصيف معاني الكلمات العربية، كـ(شجرة مفاهيم): تصنيف المعاني كـ(مفاهيم) وتحديد صفاتها الجوهرية المميزة وتمثيلها بغلة المنطق.

منهجية جديدة لتحديد وتوصيف الدلالة



https://ontology.birzeit.edu/concept/293198

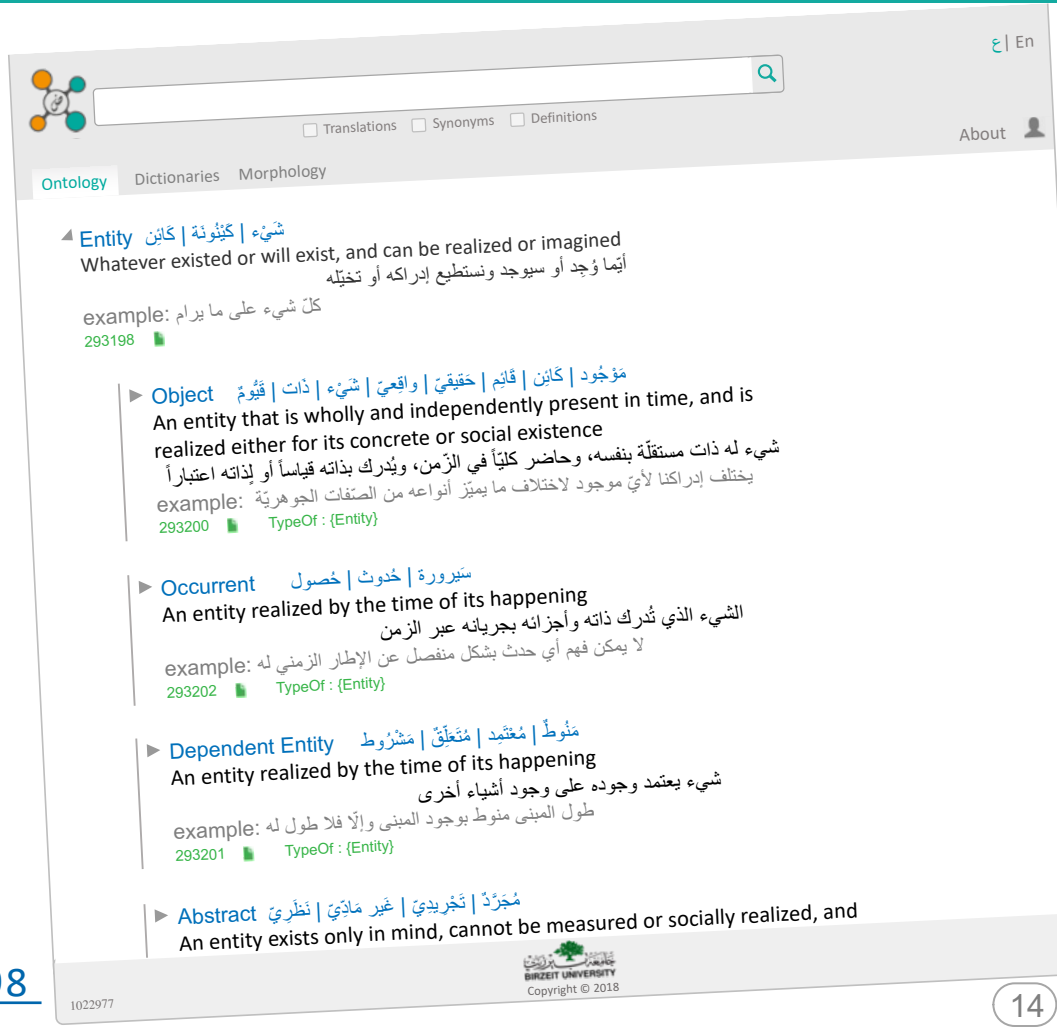
Arabic Ontology

- Each concept is a class of instances
- Each node in the ontology tree is either concept or instance.
- Each concept is described by a gloss, example, and synonyms, etc.
- Formal Arabic Wordnet - with ontologically-clean content. The Arabic Ontology is **the next generation of a wordnet**.

➢ Linked with WordNet, Wikidata,

BFO, DOLCE,

Qabas, and many lexicons



https://ontology.birzeit.edu/concept/293198

Arabic Ontology

- Current size so far (but the numbers are dynamic)

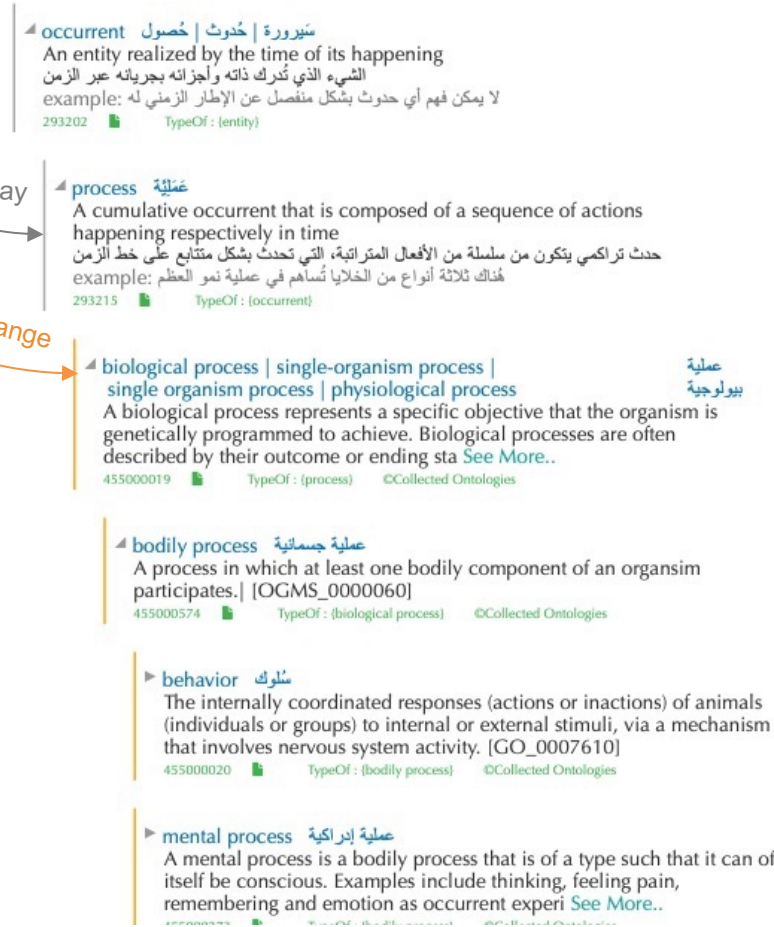
  1800 fully-done concepts (mostly top levels)

  17K partially investigated (good for NLP)

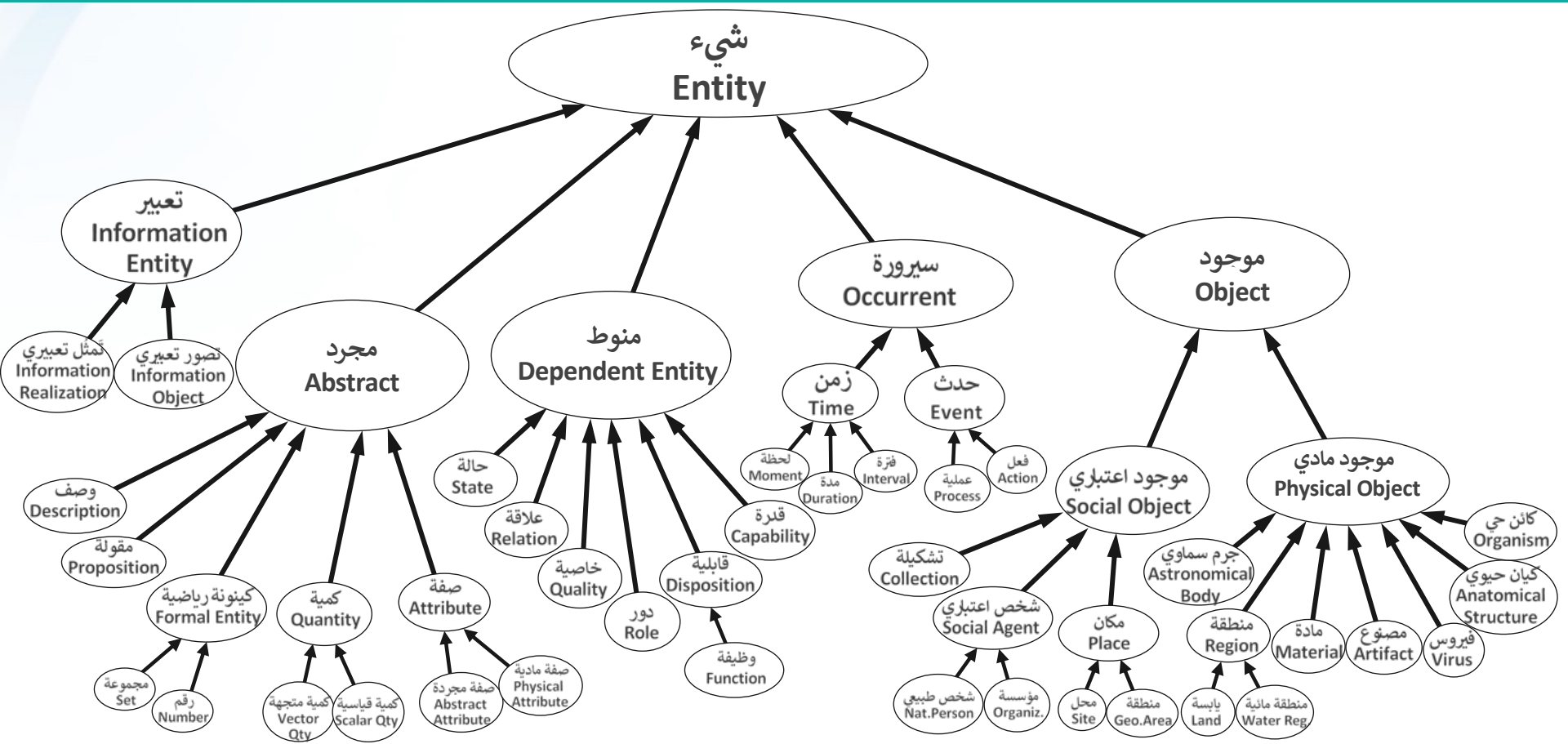
  Some branches are elaborated, other not yet.

- We provide English labels sometimes (but not always) for readability and communication.

- Methodology: built **manually** top-down and bottom-up at the same time.



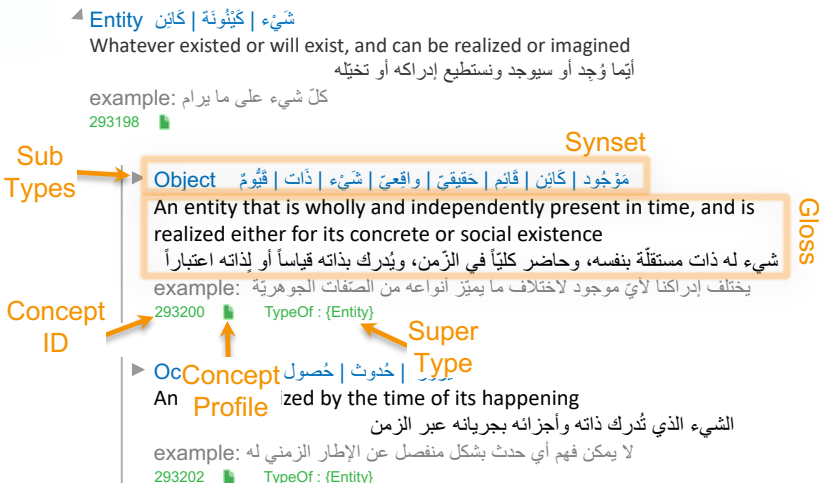
Top Levels of the Arabic Ontology



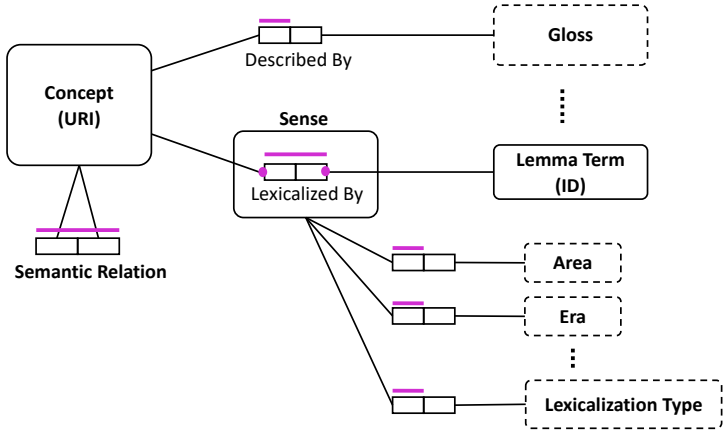
Data Representation

The general structure (i.e. core data model) of the Arabic Ontology is compatible with to the structure of WordNet – So to help in concept-mapping and interoperability.

Ontology Portal

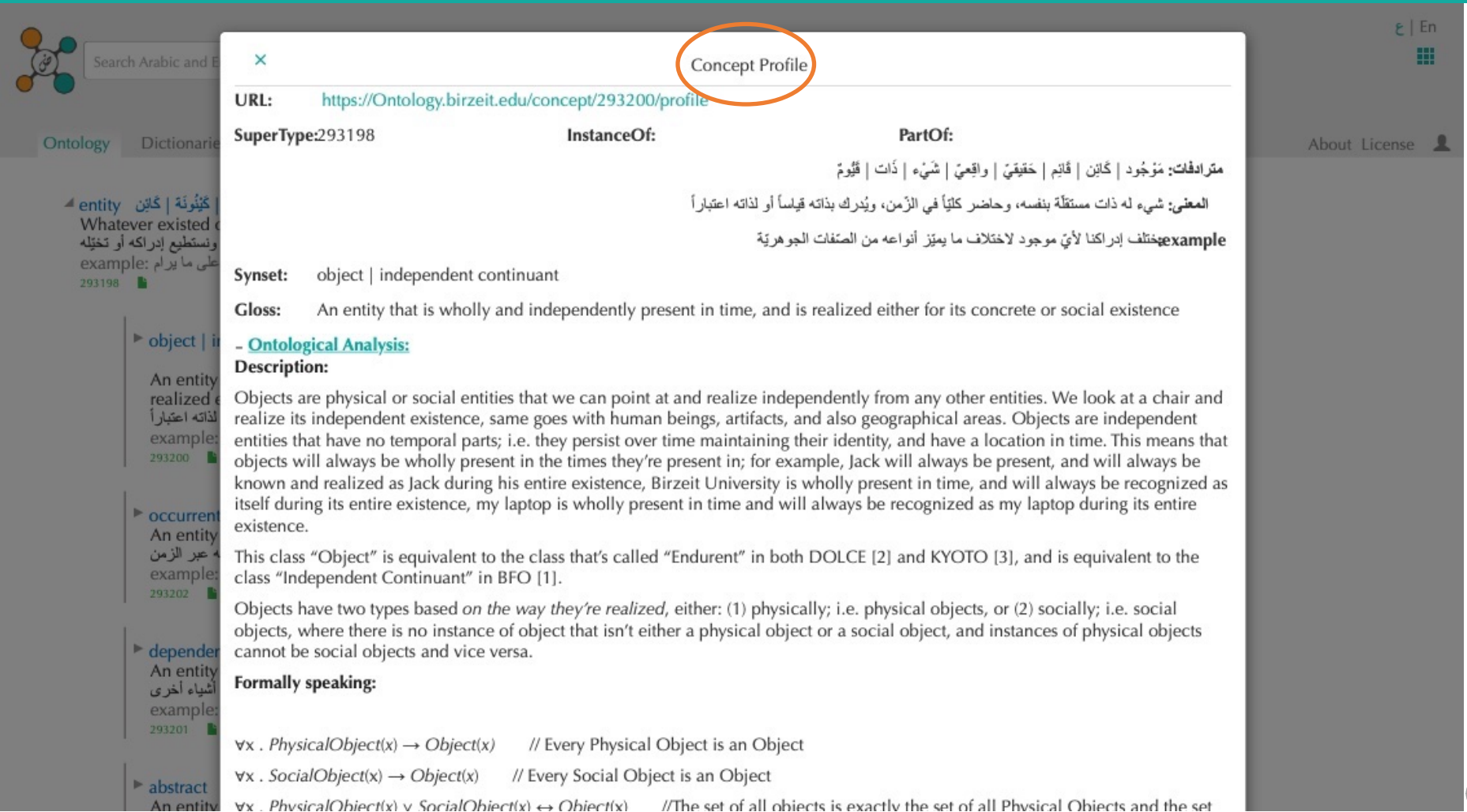


Meta Core Data Model



➔ Also accessible in the RDF W3C Lemon format

Concept Profile and Formalisms



Concept Profile

URL: https://Ontology.birzeit.edu/concept/293200/profile

SuperType:293198 **InstanceOf:** **PartOf:**

مترادفات: مَوْجُود | كَائِن | قَائِم | حَقِيقِيّ | واقِعِيّ | شَيْء | ذَات | قَيُّومْ

المعنى: شيء له ذات مستقلّة بنفسه، وحاضر كلياً في الزّمن، ويُدرك بذاته قياساً أو لذاته اعتباراً

exampleيختلف إدراكنا لأيّ موجود لاختلاف ما يميّز أنواعه من الصّفات الجوهريّة

Synset: object | independent continuant

Gloss: An entity that is wholly and independently present in time, and is realized either for its concrete or social existence

- **Ontological Analysis:**

Description:

Objects are physical or social entities that we can point at and realize independently from any other entities. We look at a chair and realize its independent existence, same goes with human beings, artifacts, and also geographical areas. Objects are independent entities that have no temporal parts; i.e. they persist over time maintaining their identity, and have a location in time. This means that objects will always be wholly present in the times they're present in; for example, Jack will always be present, and will always be known and realized as Jack during his entire existence, Birzeit University is wholly present in time, and will always be recognized as itself during its entire existence, my laptop is wholly present in time and will always be recognized as my laptop during its entire existence.

This class "Object" is equivalent to the class that's called "Endurent" in both DOLCE [2] and KYOTO [3], and is equivalent to the class "Independent Continuant" in BFO [1].

Objects have two types based *on the way they're realized*, either: (1) physically; i.e. physical objects, or (2) socially; i.e. social objects, where there is no instance of object that isn't either a physical object or a social object, and instances of physical objects cannot be social objects and vice versa.

Formally speaking:

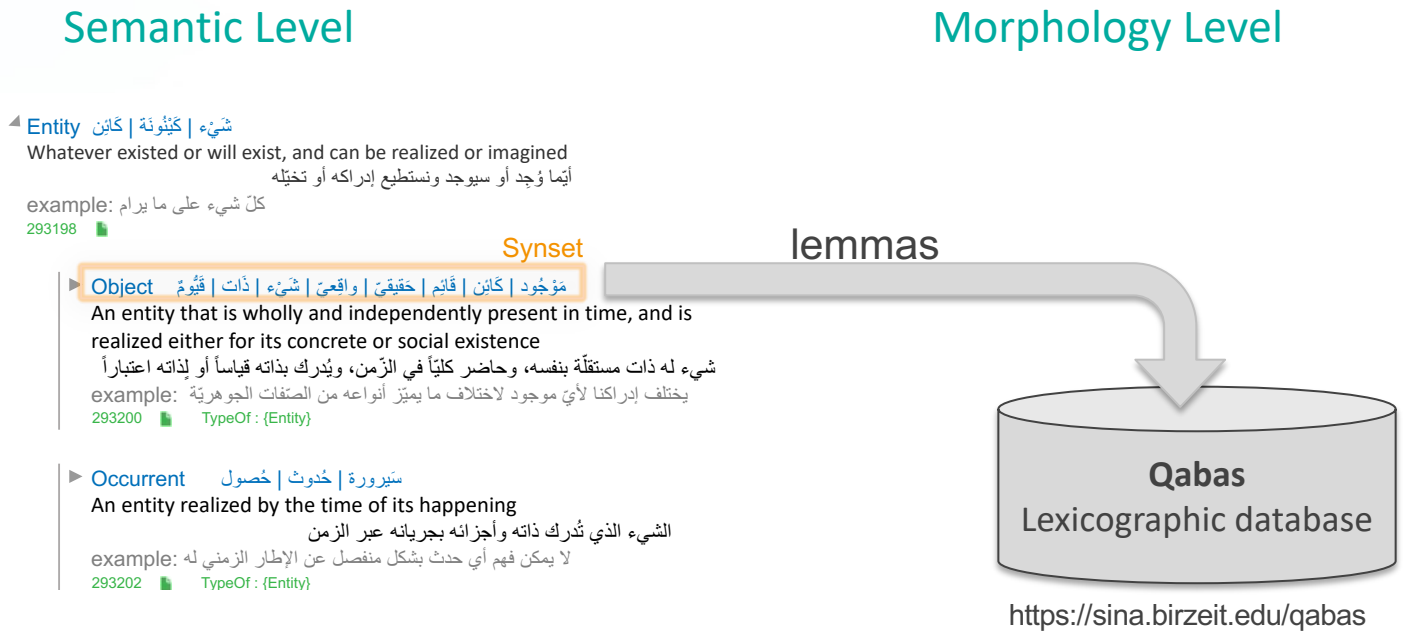
$\forall x\ .\ PhysicalObject(x) \rightarrow Object(x)$ // Every Physical Object is an Object

$\forall x\ .\ SocialObject(x) \rightarrow Object(x)$ // Every Social Object is an Object

$\forall x\ .\ PhysicalObject(x) \vee SocialObject(x) \leftrightarrow Object(x)$ //The set of all objects is exactly the set of all Physical Objects and the set

Morphology Level

All terms in the synsets are lemmas in the Qabas (Morphology Database).
Thus, linking the semantic level with the morphology level.



URIs Design

To be linked with other resources in **the Linguistic Linked Open Data Cloud**, the URLs are designed according to the **W3C's Best Practices for Publishing Linked Data**, as the following:

***Concepts*:** each concept is given a URL based on its unique ConceptID:

`https://{domain}/concept/{ConceptID}`

e.g., https://ontology.birzeit.edu/concept/293254

Semantic relations: to allow one to retrieve, e.g. the instances, of a given conceptID, the semantic relations for a given concept can be accessed through URLs:

`http://{domain}/concept/{Relation}/{ConceptID}`

e.g., https://ontology.birzeit.edu/concept/instances/293121
https://ontology.birzeit.edu/concept/parts/293121

***Terms*:** each term is given a URL, which refers to the set of concepts that are lexicalized using a certain term, i.e., that have this *term* among their synsets:

`https://{domain}/concept/{term}`

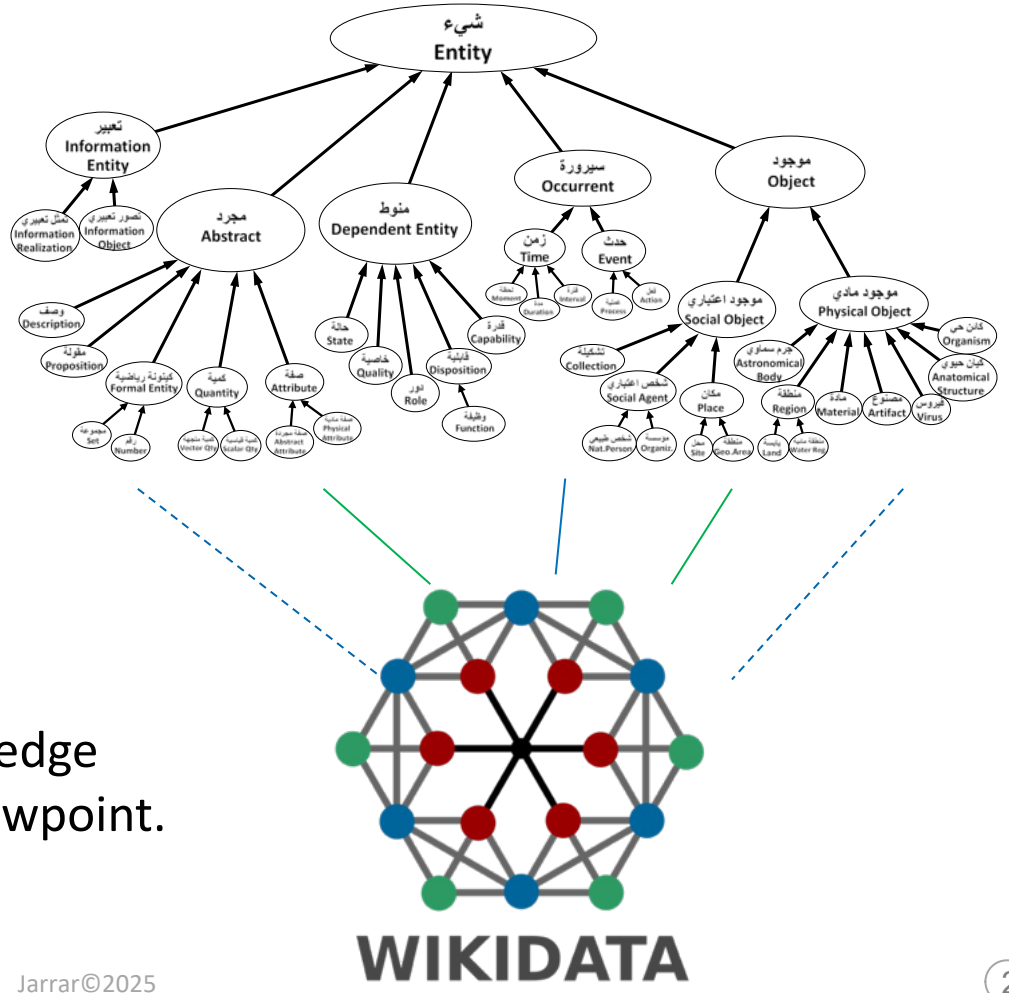
e.g., https://ontology.birzeit.edu/concept/virus

Linking with Wikidata

Every concept in the ontology is mapped with a Wikidata node

Thus, instances in Wikidata are instances in the ontology

Thus, querying the Wikidata knowledge graph from the Arabic Ontology viewpoint.



Download Arabic Ontology

Open-Source (CC-BY- 4.0)

Download: https://ontology.birzeit.edu/about

Cocnepts.csv

conceptId	arabicSynset	englishSynset	gloss	example	dataSourceId
293321	خيوط متوسطة	Intermediate Filament	تتكسب الخيوط المتوسطة الخلية الث تركيبة هيكلية خلوية تتكون من بروتين الكيراتين و تكسد		200
293572	فَتْرَةٌ	Interval | Time Interval	اللَّيل هو فترة زمنيّة بين غروب الشّ زمن يُكمَّم ويُحسب بناءً على ما يناط به من أحداث فلكيّة		200
293424	حيوان لافقاري	Invertebrate	تعج البيئة البحرية بحيوانات لافقار حيوان لا يوجد له عمود فقري		200
293243	عَدَدٌ غَيْرُ نِسْبِي | عَدَدٌ غَيْرُ كَسْرِي	Irrational Number	العدد ط هو عدد غير نسبي ، الجذو عَدَدٌ حَقِيقِيٌّ تَمْثِيلُهُ العَشْرِي غَيْرُ مُنْتَهٍ وَغَيْرُ دَوْرِي		200
48545	تسعيرة المياه | تسعير المياه | تسعيرة	irrigation rate | water a	قامت بلدية رام الله بتوحيد _ في كا ما يدفعه المنتفع مقابل إمداده بالماء للري أو لأغراض أخ		43
49035	الأسهم المباعة | الأسهم المحررة | الا	issued stock	انتهى تامر من مراجعة ملف ال_ صكّ مالي وهمي في أسواق البورصة يحرر أو يصدر ،		43
48939	حقب الحياة الحديثة | حقب السينوزوء	kainozoic era | cainozo	وجدت التدبيات منذ _ الزمن الجيولوجي الذي تلا حقب الحياة المتوسطة. ويتميز		43
169943	مَطْبَخٌ	kitchen	في دراسة تبين ان أكثر من ثلثي بو مشغل مخصص لممارسة عملية طبخ وإعداد الطعام		200
53235	مَخْبَرٌ | مُخْتَبَرٌ | مَعْمَلٌ	Lab | Laboratory	ذَهبَ إلى ال_ لعمل فحص لمعرفة معمل تُجرَى فيه التحليلات الكيماوية أو الاختبارات العلم		200
293484	يَابِسَةٌ	land	تشكل اليابسة ما يقارب ربع مساحة منطقة غير مائية من سطح الكرة الارضية		200
48095	صعوبات التعلم | صعوبات التعلم الخ	ld | learning disability |	يعاني كثير من الطلاب من _ افتقار التلميذ إلى الإنجاز أو القدرة عليه في مجال واحد أ		43
293110	مَكْتَبَةٌ	Library	تعد مكتبة بلدية البيرة أكبر مكتبة فم ملتقى مخصص لدراسة او اعارة بعض المطبوعات والم		200
53970	فحم خشبي | فحم لجنيتي | لجنيت	lignite | lignitic coal	يمتاز ال_ بلونه البني المائل للأسو فحم له مظهر الخشب، ويتكون في المرحلة التالية للحث،		43
48050	خط الانحدار الهيدرولي | خط التخلل	line of saturation | hyd	يُستخدَم _ لمعرفة منسوب المياه الـ الخط الذي يقطع الضفتين على جانبي الترعة فيرسم الحد		43
52357	سَائِلٌ 2 | رَخْوٌ | رُخْوٌ | رِخْوٌ | مَائِعٌ	liquid	يذوب الثلج ويصبح _ عند تعرضه حالة المادة بين الصلابة والغازية يتغير شكلها تبعا للوعا		43
293345	جسم مُحَلِّل	Lysosome	يعمل الجسم حال على هضم الخلاي عضية محاطة بغشاء تنشأ من جهاز غولجي، تحتوي على		200
53424	لُطْخَةٌ | بُقْعَةٌ2	macula	لاحظت رباب وجود _ على ثوبها مساحة صغيرة من سطح يخالف لونها لون بقية أجزائه.		43
54584	مَاءٌ صُهَارِيٌّ | مَاءٌ وَلِيد	magmatic water | juve	عادة ما يكون ال_ موجود تحت الأ مياه صَهيرية تخرجُ الى سَطح الارض مع مَقذوفات البر		43
51678	مَغْنِيتِيت | مَغْنَتِيت	magnetic iron ore | ma	يستخدم _ في صناعة المغناطيسات مادة حديدية صلبة ذات شكل معين لها قطبان وتجذب نو		43
50439	سوء التكيف | سوء السلوك التكيفي	maladaptive behavior	يواجه البعض مشكلة _ مع المحيط وحالة من عدم القدرة على التأقلم مع المحيط		43

Relations.csv

concept_id	subTypeOfID	partOfID	instanceOfID
51090	293183	293121	
200720			293508
203214	293566		
213592	293171		
213689	293188		
217416	293185		
219050	293188		
222617	293175		
223866	293188		
293125			293121
293560			293508
293561			293508
293562			293508
293563			293508
293724			293516
293767			293750
293768			293750

Cite:

Mustafa Jarrar: **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content.** Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021

Natural Language Processing

The Arabic Ontology

In this lecture:

- ❑ Part 1: Need for Linguistic Ontologies
- ❑ Part 2: General Overview
- ❑ Part 3: **Upper-level Concepts**
- ❑ Part 4: Gloss Formulation Guidelines
- ❑ Part 5: Fundamentals and Formal Definitions
- ❑ Part 6: Benchmarking Ontology Content
- ❑ Part 7: Discussion
- ❑ Part 8: Practice

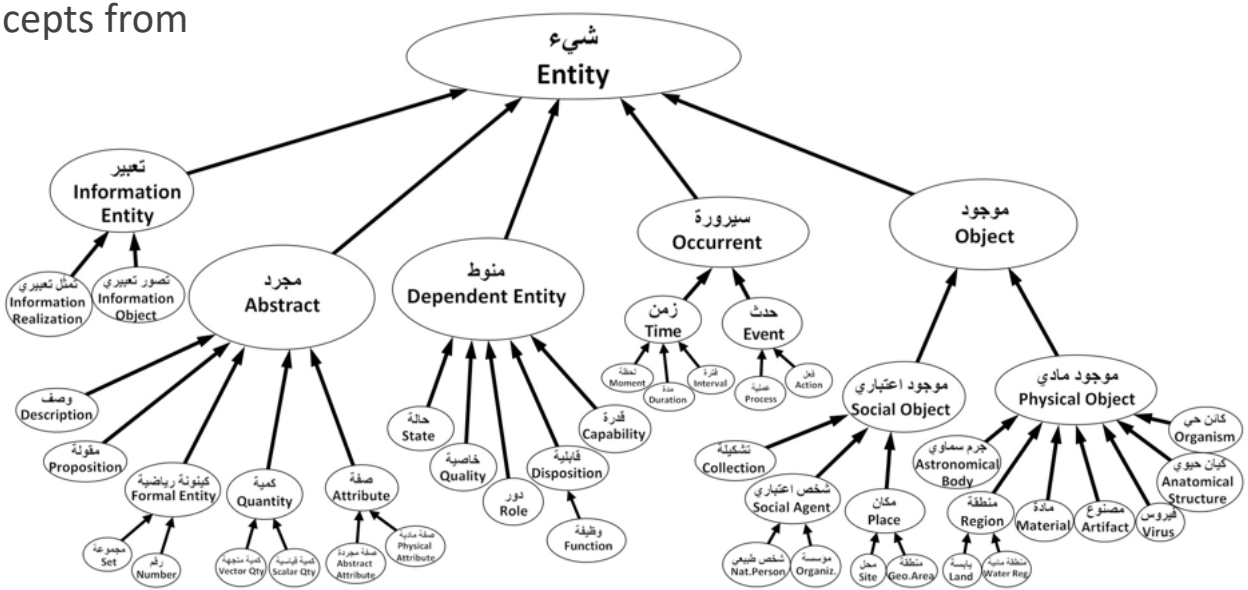
Top Levels of the Arabic Ontology

الحدود العليا - أمهات المعاني لجميع الكلمات العربية

The top levels of the Arabic Ontology tree are the most abstract concepts in Arabic; they are philosophically and logically well defined, also, linked with BFO and DOLCE upper-levels.

Why these top levels are so important:

- Derive/specialize all Arabic concepts from these top levels.
- Allows us to detect any logical and ontological mistakes in the lower levels.
- Used to governs the correctness and the evolution of the lower levels.



Top Levels of the Arabic Ontology

الحدود العليا - أمهات المعاني لجميع الكلمات العربية

The top levels of the Arabic Ontology tree are the most abstract concepts in Arabic; they are philosophically and logically well defined, also, linked with BFO and DOLCE upper-levels.

Design principles:

هل تصنيف المعاني (شجرة الانطولوجيا) جامعة مناعة؟

Totality (جامع):

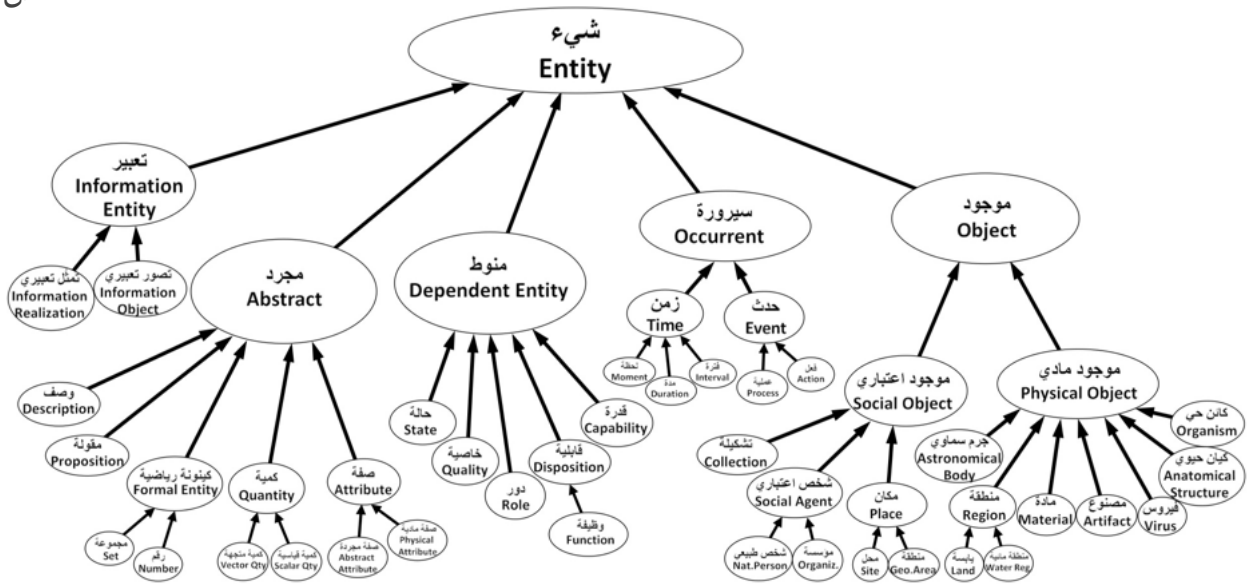
No other nodes in the same level

by: Comprehensiveness evaluation

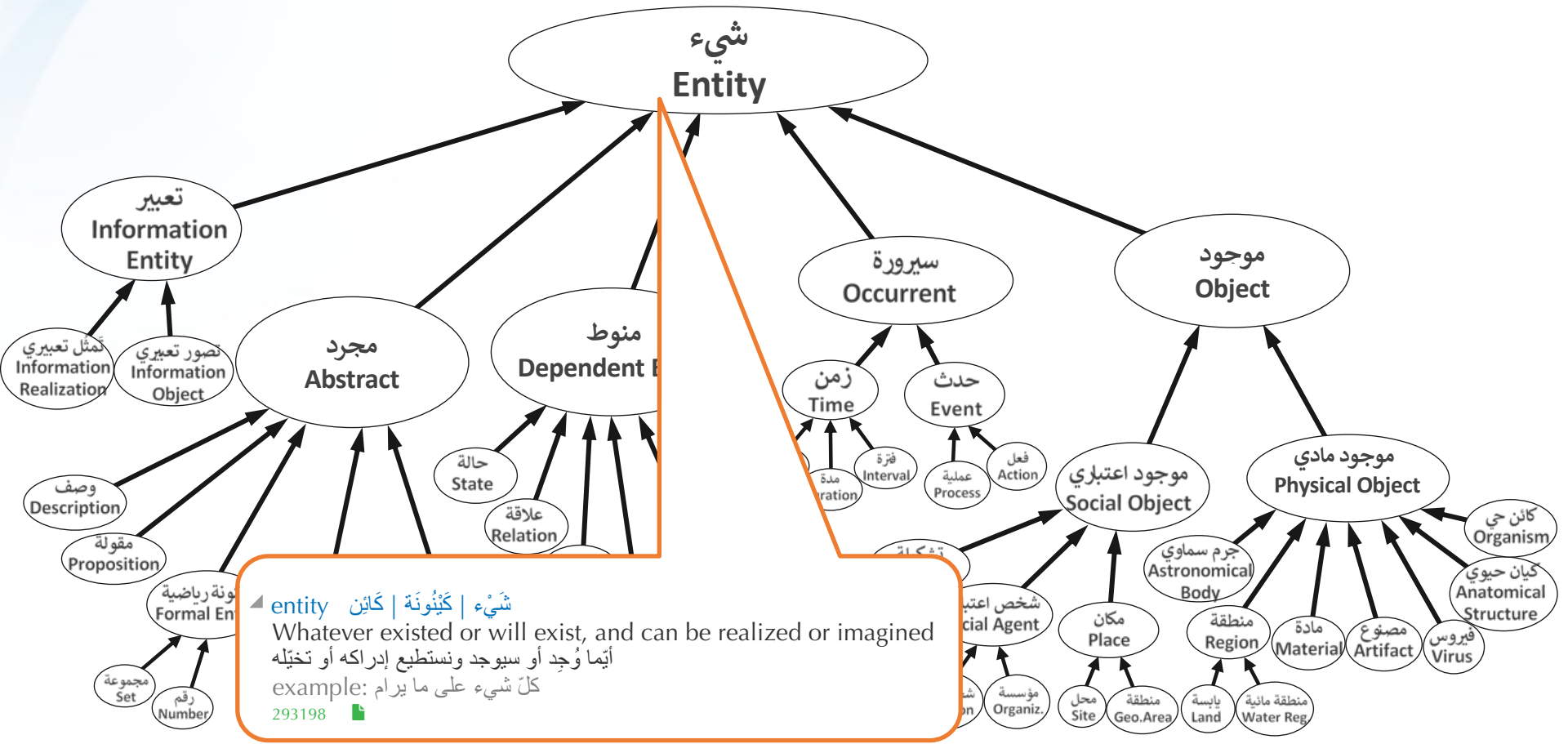
Disjointedness (مانع):

Nodes in the same level are disjoint

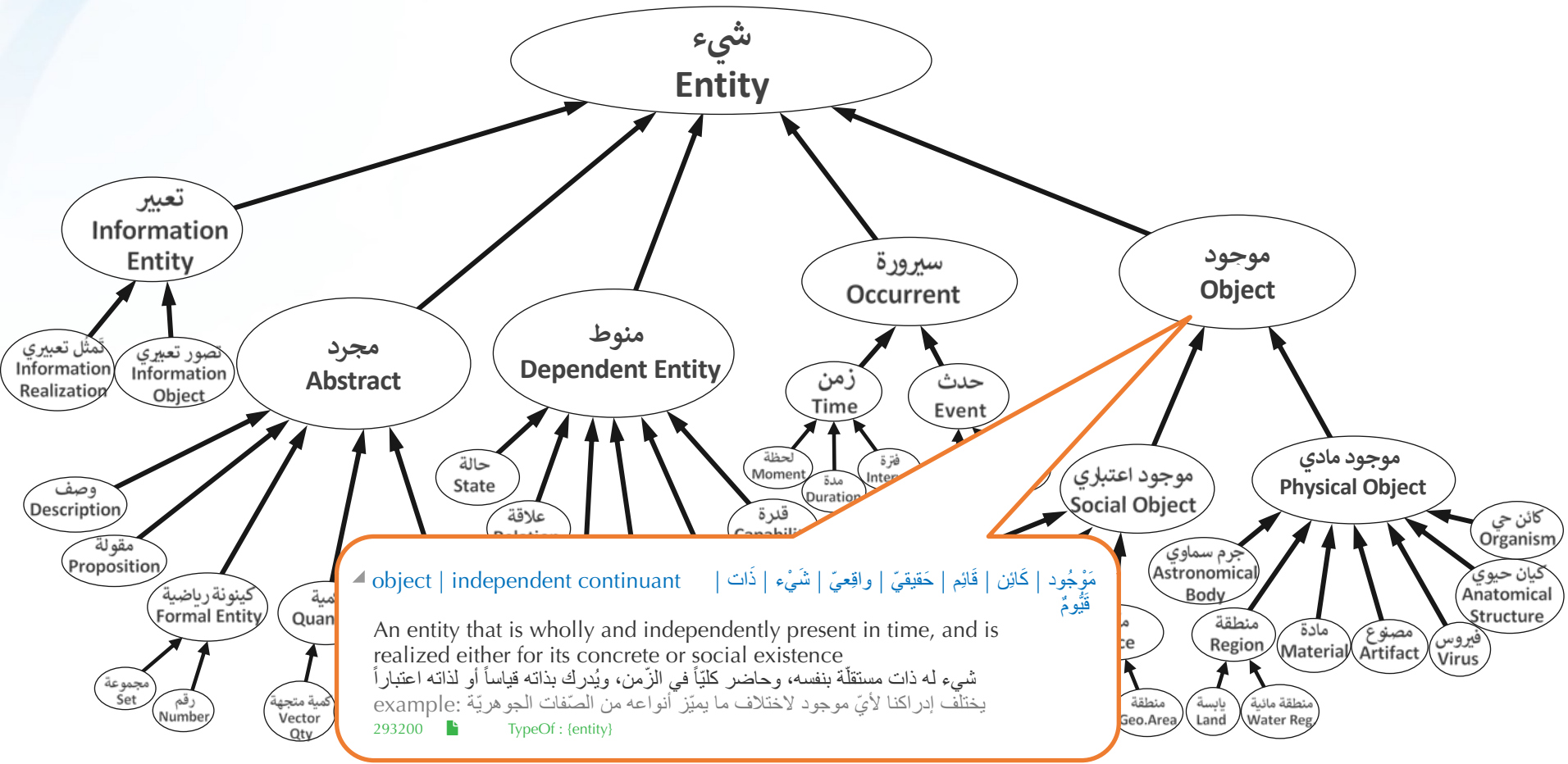
By: definition/design



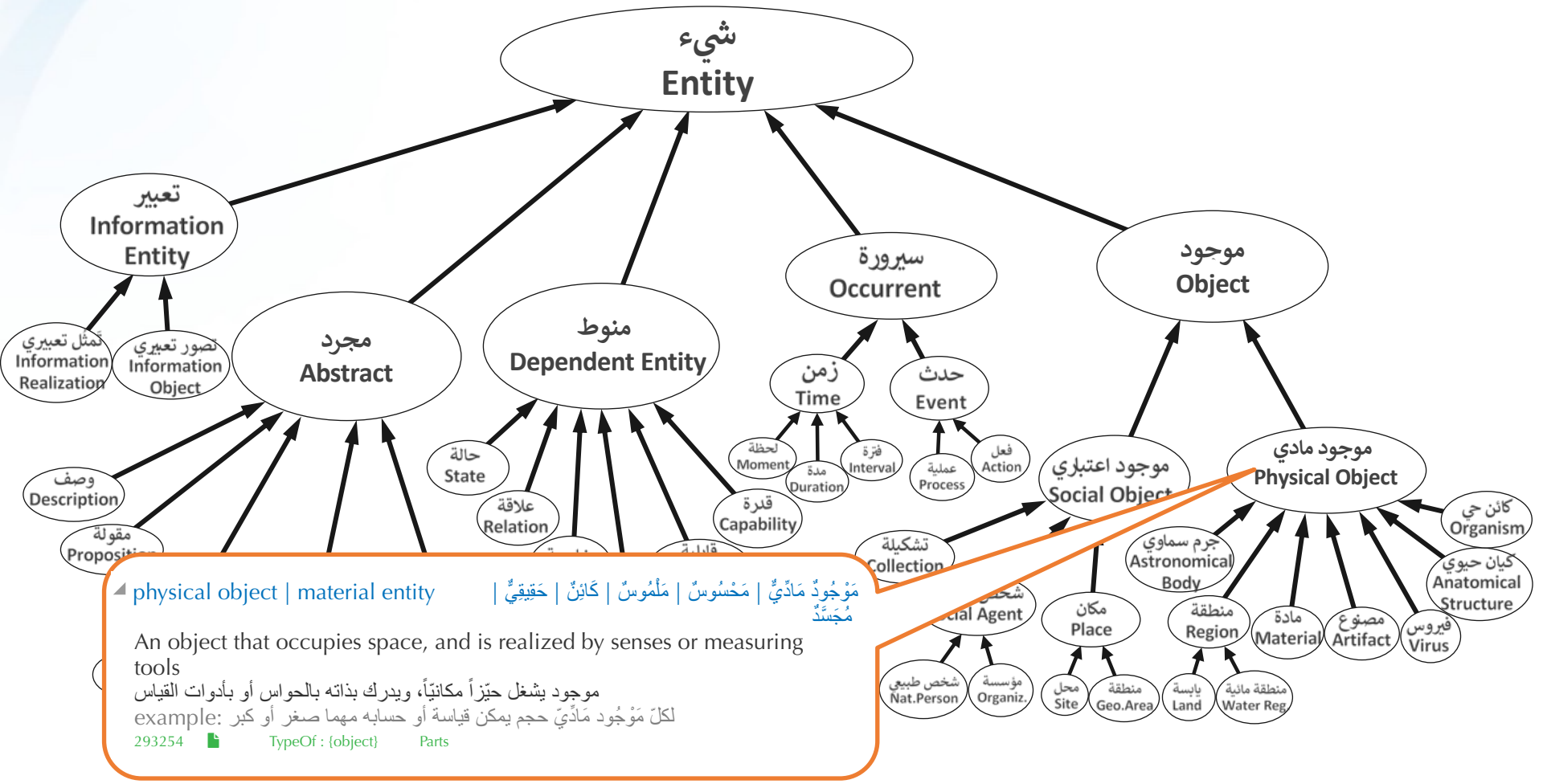
Top Levels of the Arabic Ontology



Top Levels of the Arabic Ontology



Top Levels of the Arabic Ontology



physical object | material entity | مَوْجُودٌ مَادِّيٌّ | مَحْسُوسٌ | مَلْمُوسٌ | كَائِنٌ | حَقِيقِيٌّ | مُجَسَّدٌ

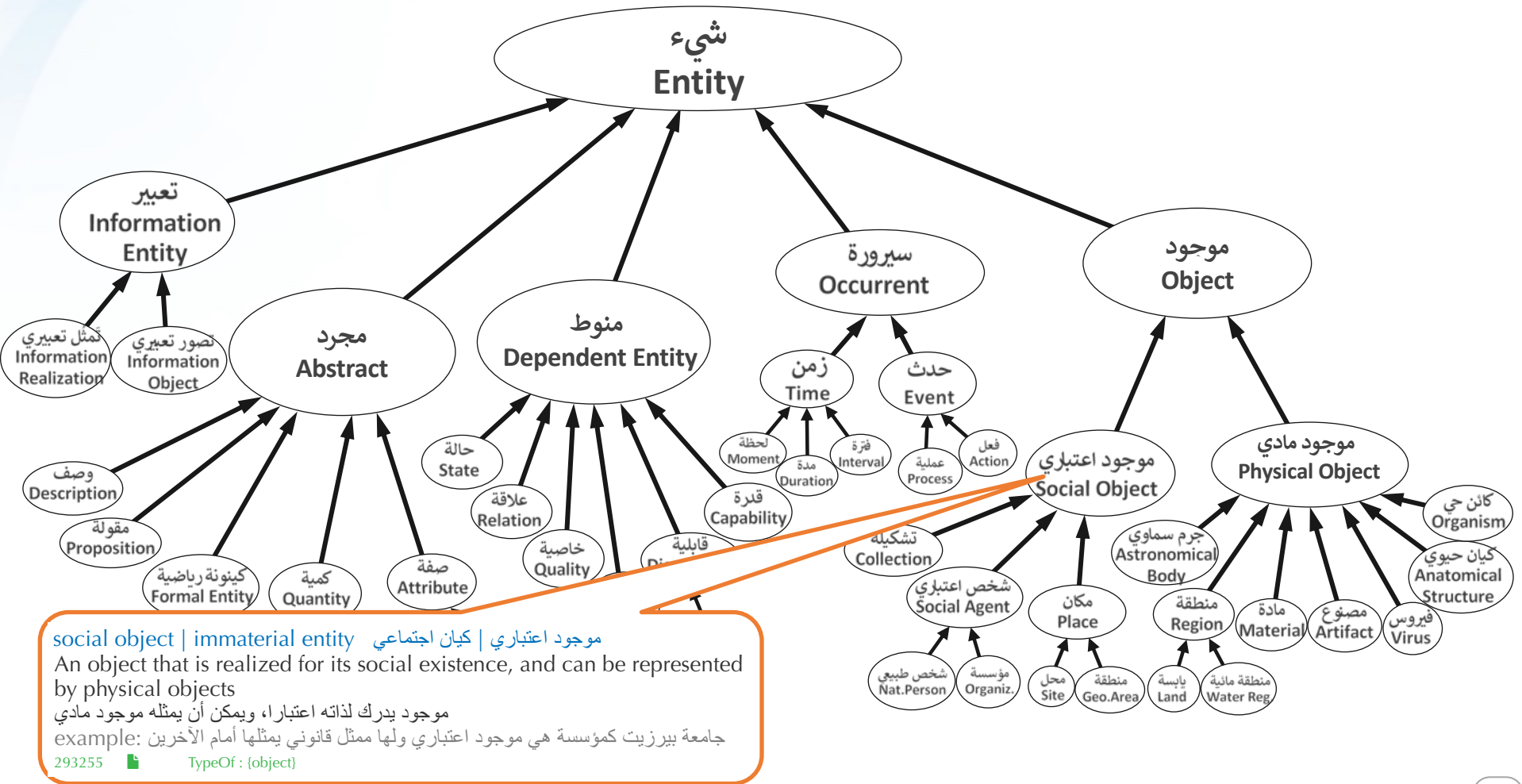
An object that occupies space, and is realized by senses or measuring tools

موجود يشغل حيّزاً مكانيّاً، ويدرك بذاته بالحواس أو بأدوات القياس

example: لكلّ مَوْجُود مَادِّيّ حجم يمكن قياسة أو حسابه مهما صغر أو كبر

293254 TypeOf : {object} Parts

Top Levels of the Arabic Ontology



social object | immaterial entity موجود اعتباري | كيان اجتماعي

An object that is realized for its social existence, and can be represented by physical objects

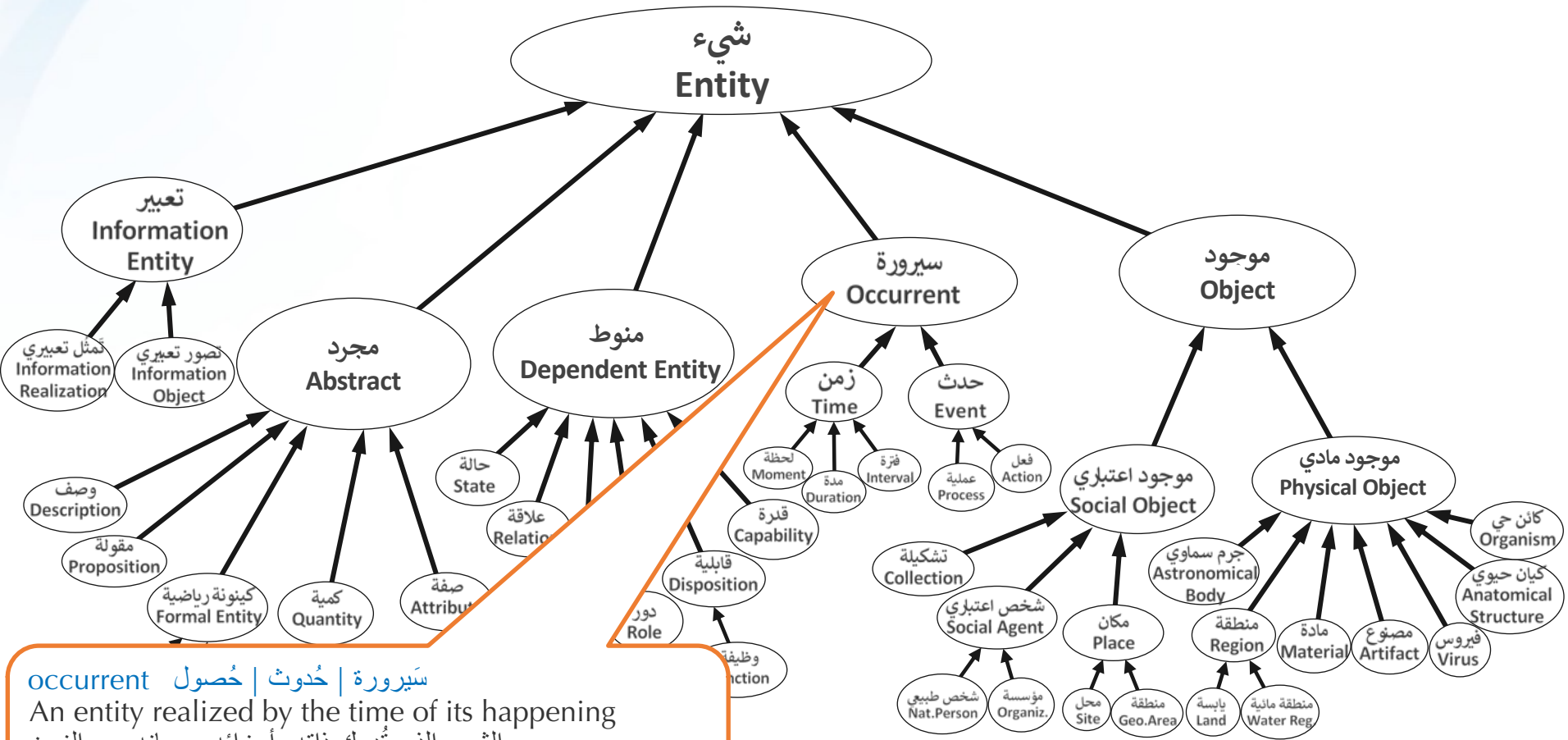
موجود يدرك لذاته اعتبارا، ويمكن أن يمثله موجود مادي

example: جامعة بيرزيت كمؤسسة هي موجود اعتباري ولها ممثل قانوني يمثلها أمام الآخرين

293255 TypeOf : {object}

Jarrar©2025

Top Levels of the Arabic Ontology



occurrent سَيرورة | حُدوث | حُصول

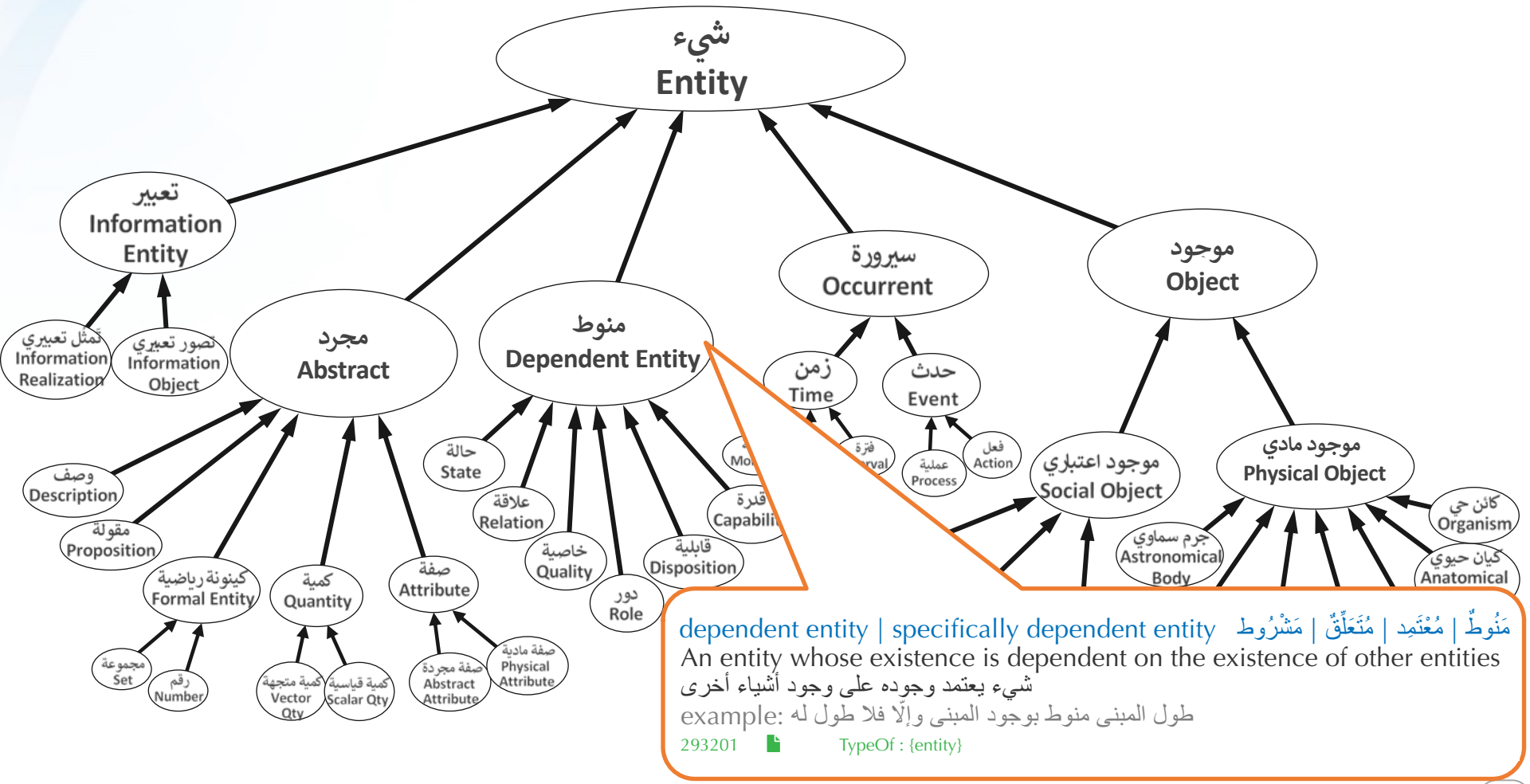
An entity realized by the time of its happening

الشيء الذي تُدرك ذاته وأجزائه بجريانه عبر الزمن

example: لا يمكن فهم أي حدوث بشكل منفصل عن الإطار الزمني له

293202 TypeOf : {entity}

Top Levels of the Arabic Ontology



dependent entity | specifically dependent entity مَنُوطٌ | مُعْتَمِد | مُتَعَلِّقٌ | مَشْرُوط

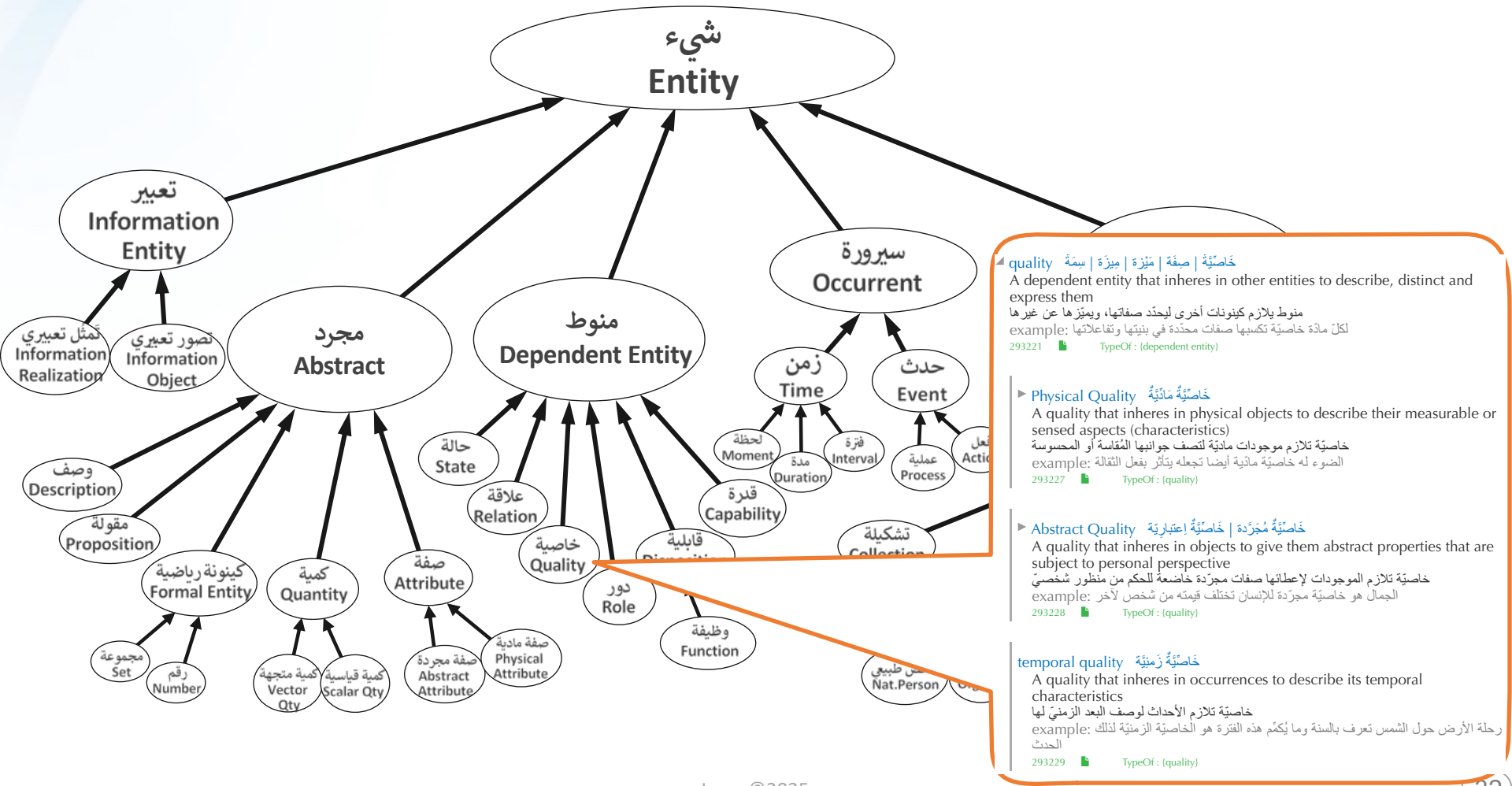
An entity whose existence is dependent on the existence of other entities

شيء يعتمد وجوده على وجود أشياء أخرى

example: طول المبنى منوط بوجود المبنى وإلّا فلا طول له

293201 TypeOf : {entity}

Top Levels of the Arabic Ontology



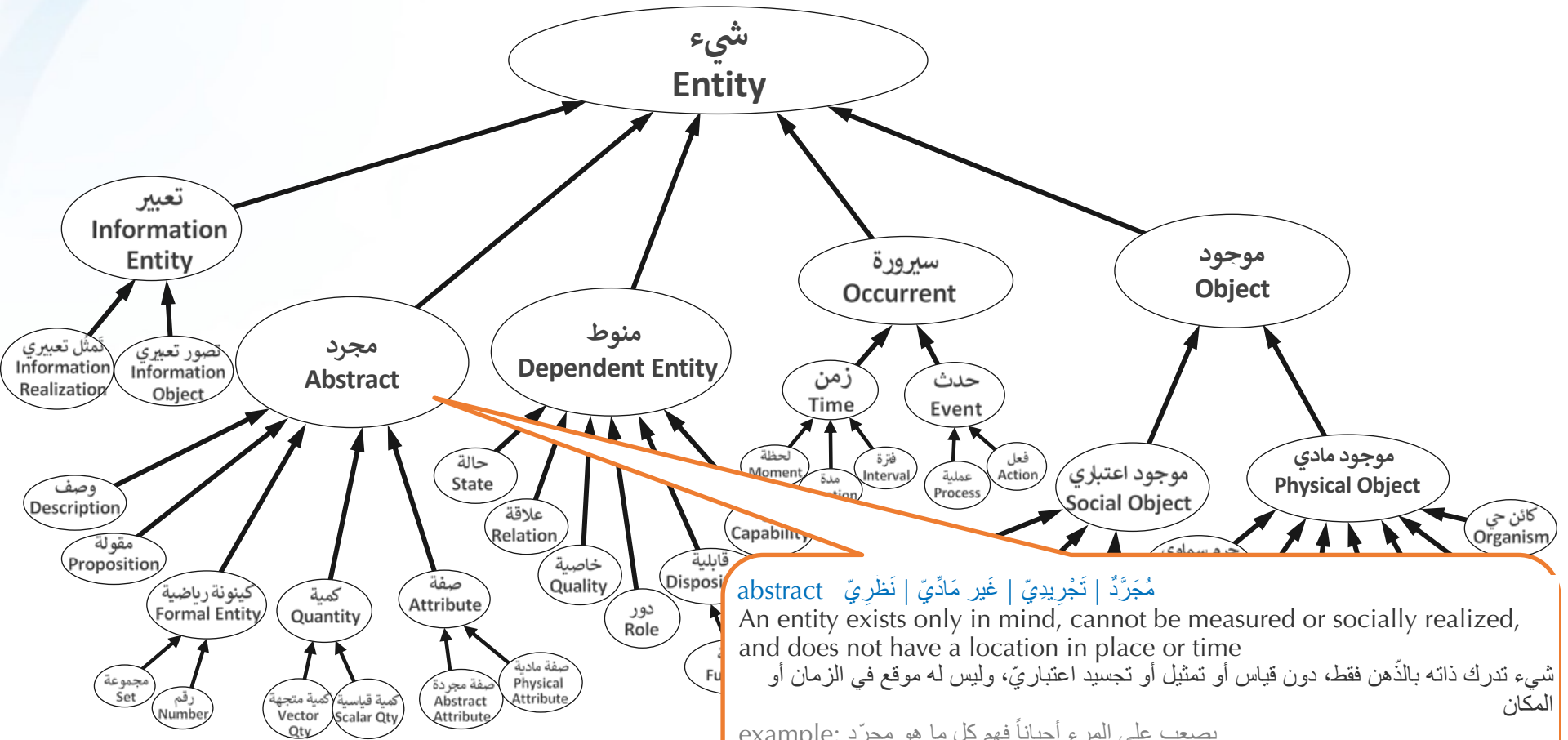
quality خَاصِّيَّة | صِفَة | مَيْزَة | مِيزَة | سِمَة
A dependent entity that inheres in other entities to describe, distinct and express them
منوط يلازم كينونات أخرى ليحدّد صفاتها، ويميّزها عن غيرها
example: لكلّ مادّة خاصيّة تكسبها صفات محدّدة في بنيتها وتفاعلاتها
293221 TypeOf : {dependent entity}

Physical Quality خَاصِّيَّةٌ مَادِّيَّةٌ
A quality that inheres in physical objects to describe their measurable or sensed aspects (characteristics)
خاصيّة تلازم موجودات ماديّة لتصف جوانبها المُقاسة أو المحسوسة
example: الضوء له خاصيّة مادّية أيضا تجعله يتأثر بفعل الثقالة
293227 TypeOf : {quality}

Abstract Quality خَاصِّيَّةٌ مُجَرَّدة | خَاصِّيَّةٌ اِعتِبارِيَّة
A quality that inheres in objects to give them abstract properties that are subject to personal perspective
خاصيّة تلازم الموجودات لإعطائها صفات مجرّدة خاضعة للحكم من منظور شخصيّ
example: الجمال هو خاصيّة مجرّدة للإنسان تختلف قيمته من شخص لآخر
293228 TypeOf : {quality}

temporal quality خَاصِّيَّةٌ زَمنِيَّة
A quality that inheres in occurrences to describe its temporal characteristics
خاصيّة تلازم الأحداث لوصف البعد الزمنيّ لها
example: رحلة الأرض حول الشمس تعرف بالسنة وما يُكمّم هذه الفترة هو الخاصيّة الزمنيّة لذلك الحدث
293229 TypeOf : {quality}

Top Levels of the Arabic Ontology



abstract مُجَرَّدٌ | تَجْرِيدِيّ | غَير مَادِّيّ | نَظرِيّ

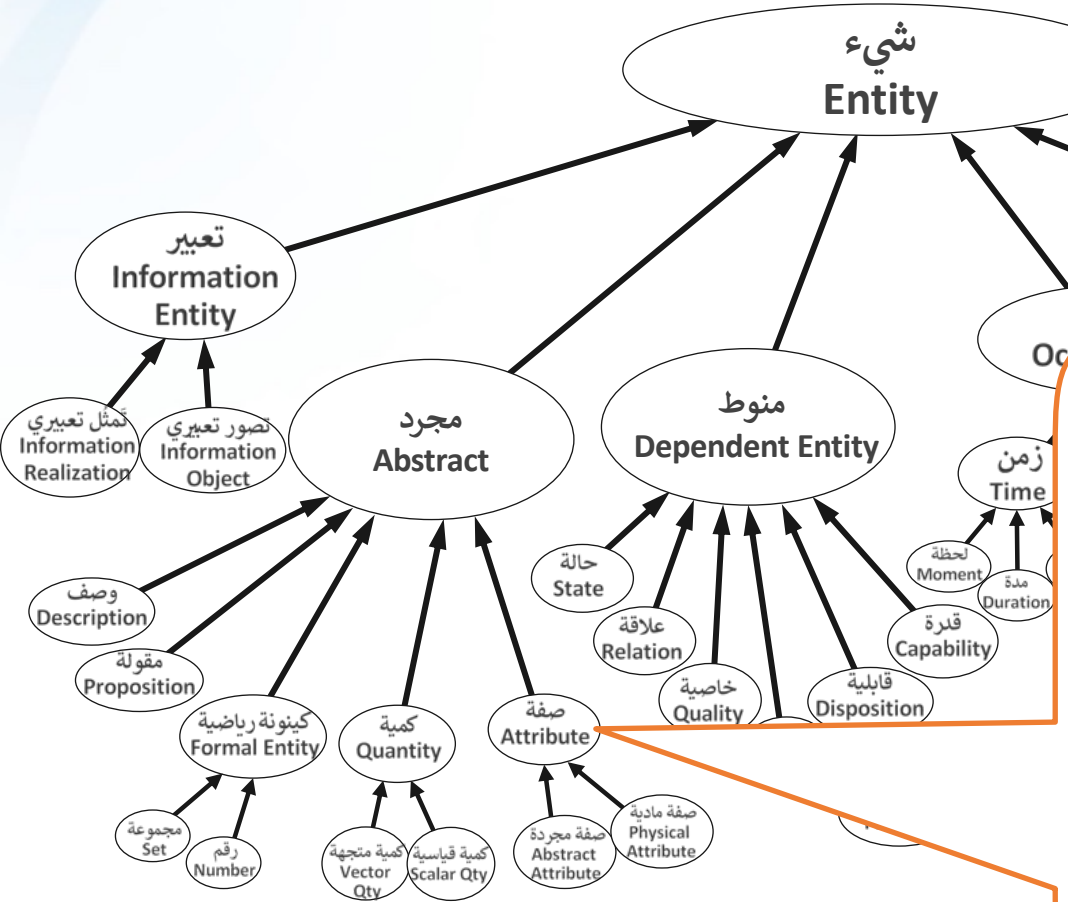
An entity exists only in mind, cannot be measured or socially realized, and does not have a location in place or time

شيء تدرك ذاته بالذّهن فقط، دون قياس أو تمثيل أو تجسيد اعتباريّ، وليس له موقع في الزمان أو المكان

example: يصعب على المرء أحياناً فهم كل ما هو مجرّد

293232 TypeOf : {entity}

Top Levels of the Arabic Ontology



attribute صِفَةٌ | خَصْلَةٌ | مُسْنَدٌ | حالَةٌ | نَعْتٌ | مَزِيَّةٌ | مَيْزَةٌ | شَكْلٌ | هَيْئَةٌ | سِمَةٌ

An abstract that describes qualities of an entity without using units o measure

مجرّد يصف ويُعبّر عن قيمة خاصيّة لشيء ما دون استعمال وحدة قياس

example: اللون البرتقاليّ هو صفة لخاصيّة اللون في البرتقال

293263 TypeOf : {abstract}

▶ Physical Attribute صِفَةٌ مَادِّيَّةٌ

An attribute that describes the value of a physical quality

صفة للتعبير عن قيمة خاصيّة مادّيّة

example: الأخضر هي صفة مادّيّة لخاصيّة لون عيون بعض الأشخاص

293265 TypeOf : {attribute}

▶ Abstract Attribute صِفَةٌ مُجَرَّدةٌ | صِفَةٌ اِعْتِبارِيَّةٌ

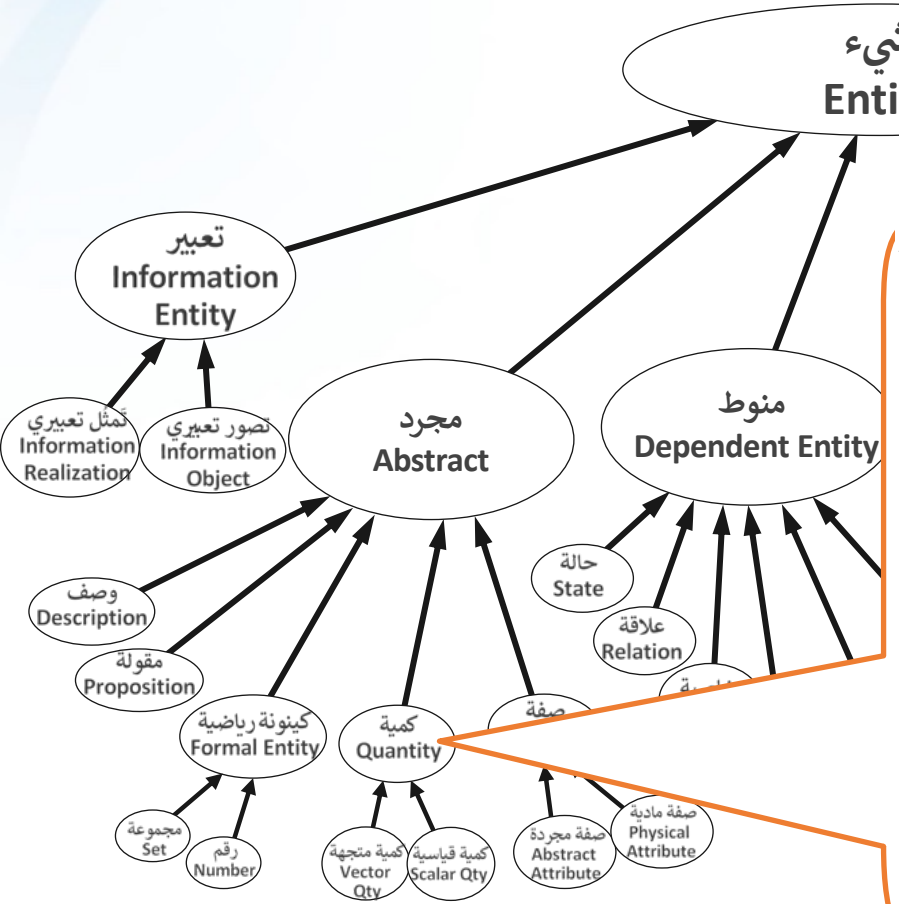
Attribute that describes an abstract quality of things

صفة تعبّر عن قيمة خاصيّة مجرّدة لشيء ما

example: كريم هي صفة مجرّدة لخاصيّة العطاء عند الإنسان

293264 TypeOf : {attribute}

Top Levels of the Arabic Ontology



quantity كَمِّيَّةٌ | مِقْدَارٌ | مِقْيَاسٌ | قَدْرٌ | قَدَرٌ | قِسْطٌ

An abstract that measures or quantifies the quality of an entity by units of measure

مجرّد يُعبّر عن قيمة لخاصيّة شيء ما بوحدة قياس

example: الكتلة كمّية لها وحدة عالميّة تعرف بالكيلوغرام

293266 TypeOf : {abstract}

Scalar Quantity كَمِّيَّةٌ عَدَدِيَة | كَمِّيَّةٌ غَيرَ مُتَّجِهَةٍ | كَمِّيَّةٌ قِياسِيَّةٌ | كَمِّيَّةٌ مَعْدُودَةٌ

A quantity determined only in magnitude, without direction

كميّة تُحَدّد بمقدار فقط دون اتّجاه

example: تستخدم عدّة أدوات لتحديد مقادير أي كميّة قياسيّة بحسب نوع تلك الكميّة

293267 TypeOf : {quantity}

Vector Quantity كَمِّيَّةٌ مُتَّجِهةٌ

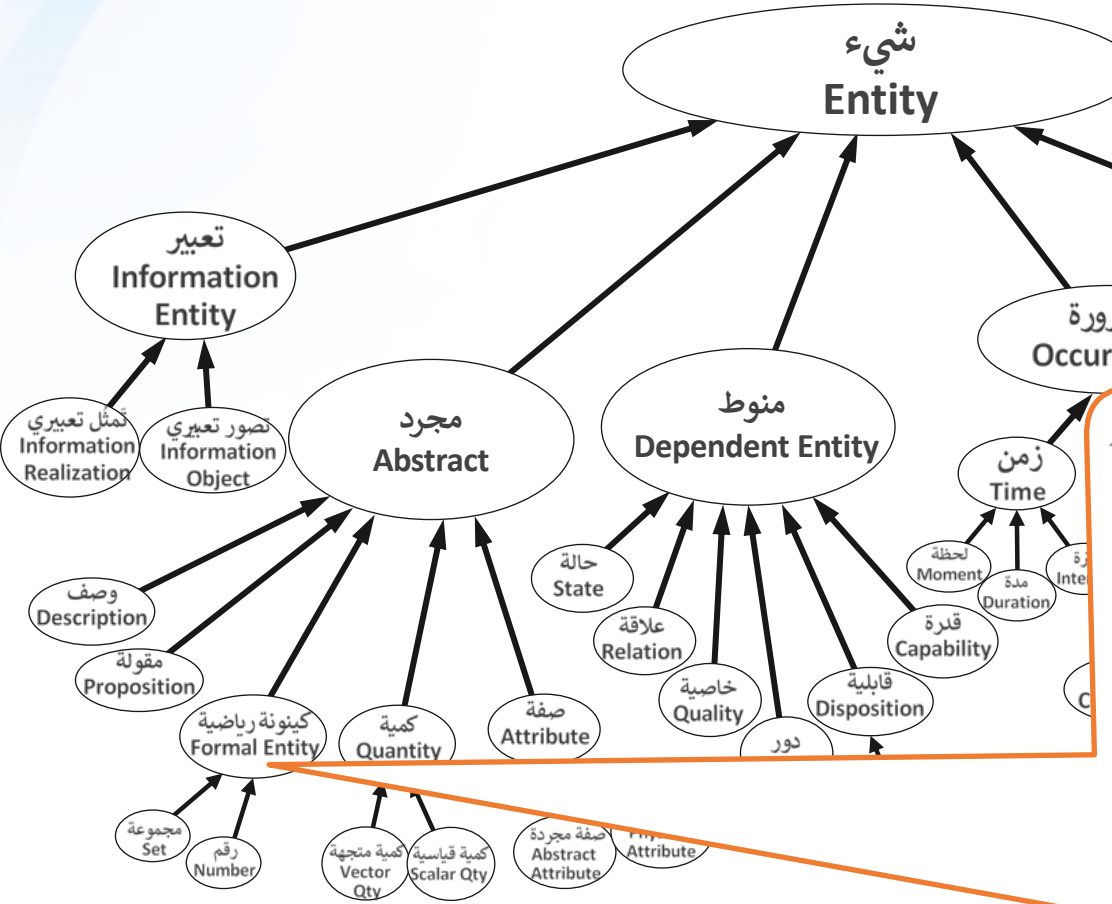
A quantity that is determined in magnitude, direction and a point of impact

كميّة تتحدّد بالمقدار والاتّجاه ونقطة التأثير

example: السرعة هي كميّة متّجهةٌ بينما مقدارها كميّة قياسيّة

293285 TypeOf : {quantity}

Top Levels of the Arabic Ontology



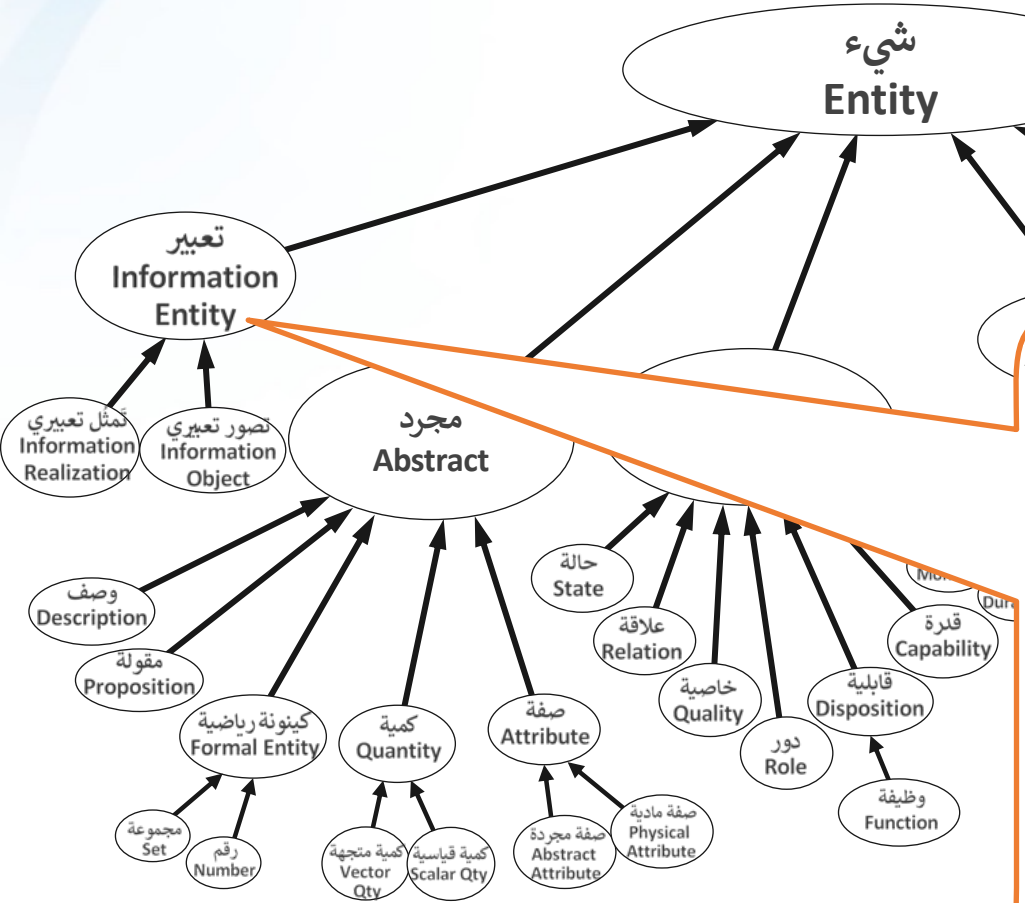
formal entity حَدٌّ رِياضِيٌّ | حَدٌّ هَنْدسيّ | كَيْنونةٌ رِياضِيّةٌ
An abstract entity that is formally defined either mathematically or logically, and does not need to be proved
شيء مجرّد اصطلح عليه رياضيّاً أو منطقيّاً، وليس بحاجة إلى برهان
example: باي هو حدّ رياضيّ
293282 TypeOf : {abstract}

number عَدَدٌ | رَقَمٌ
A formal entity representing a quantity, used in counting and measuring
حدّ رياضيّ يعبّر عن كميّة، يستعمل في العدّ والقياس
example: يمتلك قائد فريق كرة القدم العدد الحادي عشر
293235 TypeOf : {formal entity}

set مَجْمُوعَةٌ | فِئَةٌ
A formal entity that consists of a number of numbers and symbols, and is referred to as a whole unit
حدّ رياضيّ يتكوّن من أعداد أو رموز أو حدود رياضيّة أخرى، ويشار إليه كوحدة واحدة
example: كُلّ عنصر في المجموعة الجزئيّة هو عنصر في المجموعة الكليّة
293236 TypeOf : {formal entity}

Jarrar©2025

Top Levels of the Arabic Ontology



information تَعْبِير

An entity that relationally represents other entities, bearing informational content but independent from how this content is recognized

شيء يمثل أو يدلل على كينونات أخرى علائقياً، ويحمل معلومة بغض النظر عن كيفية تمثيل أو تجسيد هذه المعلومة

example: يدلل التعبير بطرق علمية مختلفة مكتوبة أو مرئية

293203 TypeOf : {entity} Parts

information object مَعْلُومَة | كِيَانٌ مَعْلُومَاتي

The information itself independent from how its content is realized

تعبير يدلل ويربط موجود مادي بمعنى لموضوع ذهني، بغض النظر عن كيفية تمثيل او تجسيد هذا التعبير في الواقع

example: سمعت بهذه المعلومة بأكثر من طريقة ومن أكثر من مصدر

293473 TypeOf : {information}

information realization تَعْبِيرٌ مُجَسَّدٌ

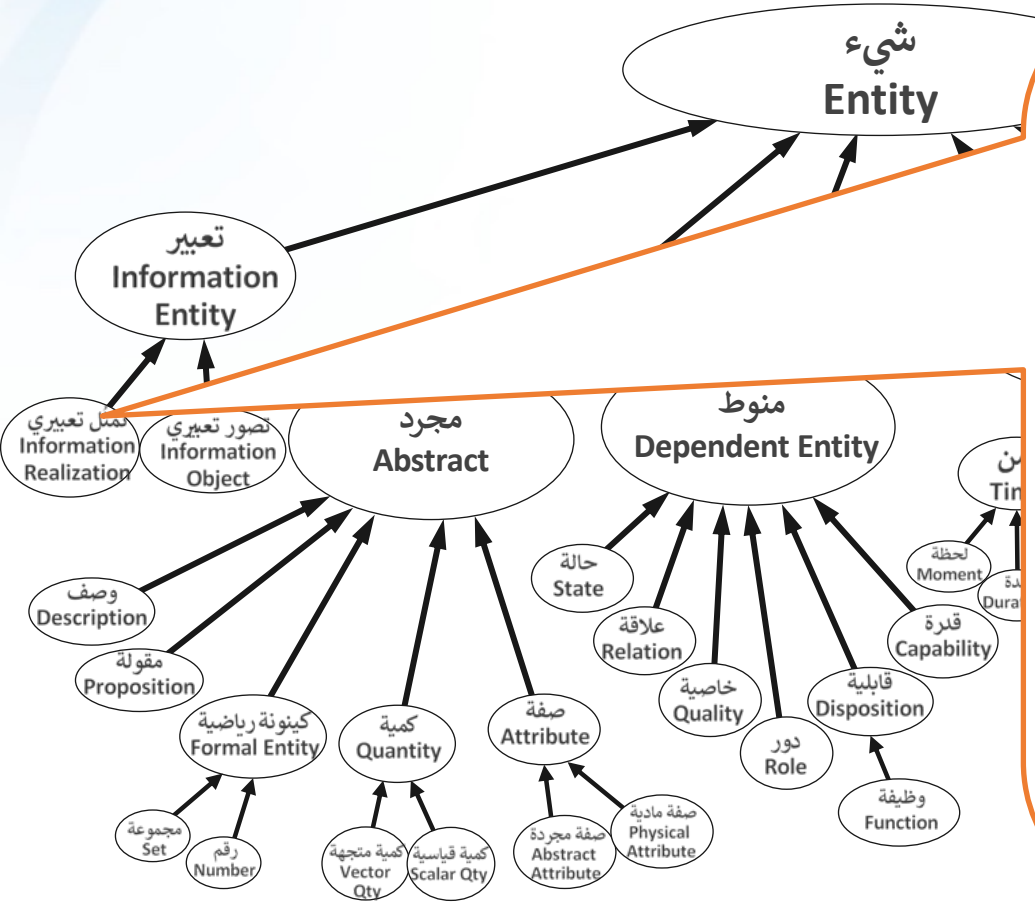
Information realizated and represented concretly

تعبير مجسد ومحقق في الواقع

example: قدم الطالب أطروحة موقعة للمكتبة وكانت تعبير مجسداً لما أنجزه خلال بحثه

293472 TypeOf : {information}

Top Levels of the Arabic Ontology



information realization تَعْبيرٌ مُجَسَّدٌ

Information realizated and represented concretly

تعبير مجسد ومحقق في الواقع

example: قدم الطالب أطروحة موقعة للمكتبة وكانت تعبير مجسداً لما أنجزه خلال بحثه

293472 TypeOf : {information}

data item | data بيانات

a data item is an information content entity that is intended to be a truthful statement about something (modulo, e.g., measurement precision or other systematic errors) and is con See More..

200000023 TypeOf : {information realization}

centrally registered identifier | CRID معرف المسجل الرئيسي

An information content entity that consists of a CRID symbol and additional information about the CRID registry to which it belongs.[IAO]

200000025 TypeOf : {information realization}

datum label عنوان

A label is a symbol that is part of some other datum and is used to either partially define the denotation of that datum or to provide a means for identifying the datum as a member See More..

200000027 TypeOf : {information realization}

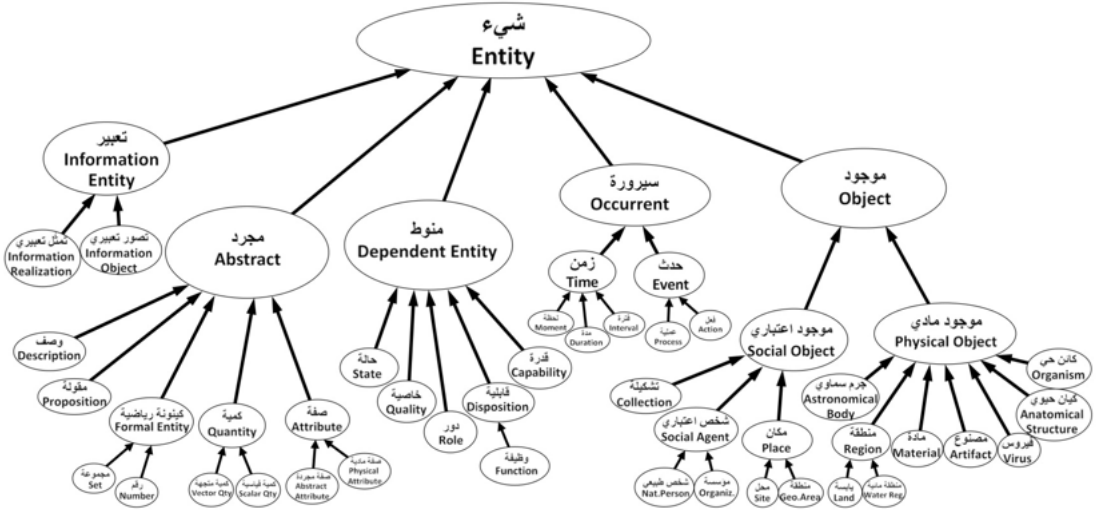
word | word form كَلِمَة | كَلِمَة

تعبير كتابي للإشارة أو الدلالة على معنى

200000034 TypeOf : {information realization}

Comprehensiveness Evaluation

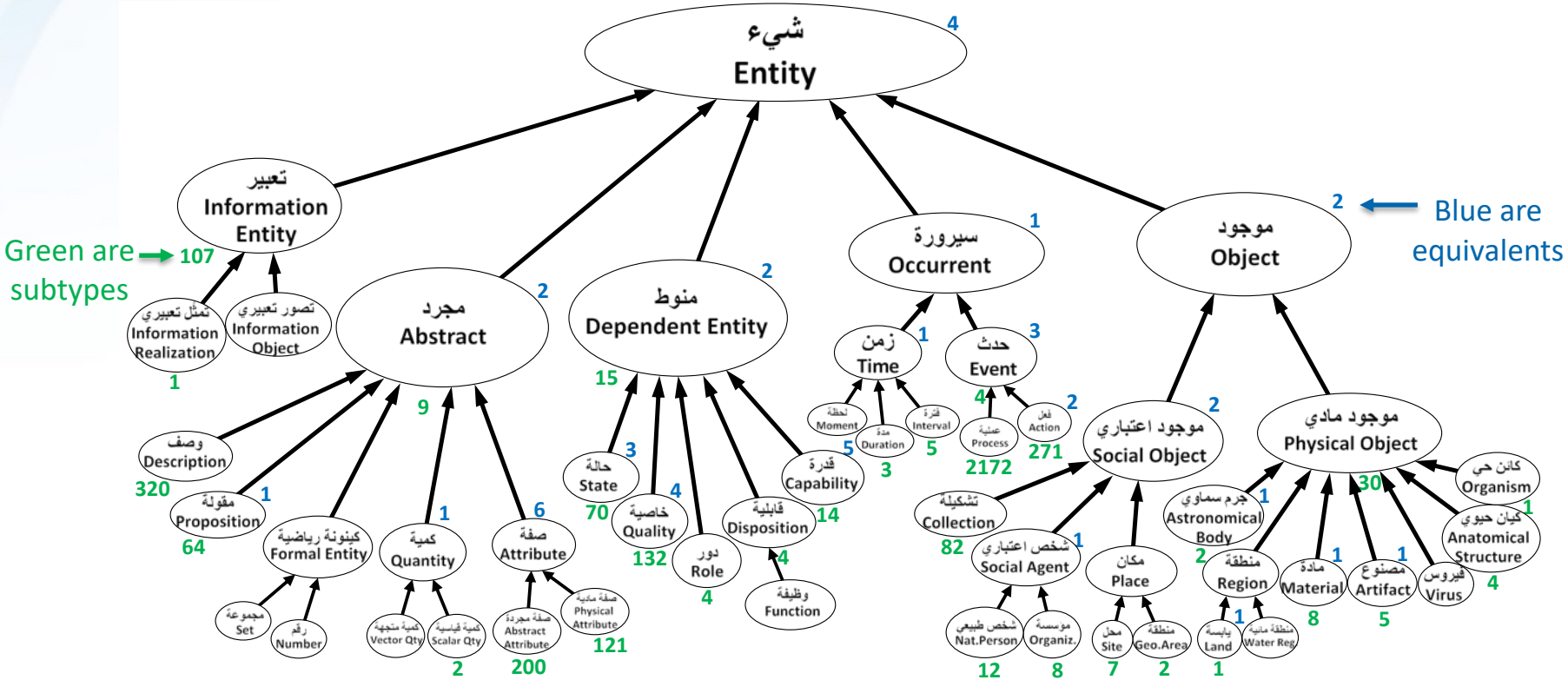
How much is the current version of the ontology is able to *top* (i.e. be a supertype of) the concepts of the Arabic terms?



Experiment: classify the 1830 terms found in Al-Jurjānī "Taʿrīfāt" lexicon (1339–1414 AD) under the ontology. This lexicon contains the most abstract notions in Arabic in most domains. 270 concepts are excluded (not understand, can be instantiated , etc).

Ideally: each of the 1830 concepts should be placed either as "equivalent to" a node in the ontology, or as a "*subtype* of" a *leaf node*.

Comprehensiveness Results



- 1655 (90%) concepts that are correctly placed in the ontology.
- 156 (10%) concepts that are mapped as subclasses of non-leaf nodes illustrate cases of missing top categories in the ontology (we are working on adding them).

Natural Language Processing

The Arabic Ontology

In this lecture:

- ❑ Part 1: Need for Linguistic Ontologies
- ❑ Part 2: General Overview
- ❑ Part 3: Upper-level Concepts
- ❑ Part 4: **Gloss Formulation Guidelines**
- ❑ Part 5: Fundamentals and Formal Definitions
- ❑ Part 6: Benchmarking Ontology Content
- ❑ Part 7: Discussion
- ❑ Part 8: Practice

Gloss Formulation Methodology

Gloss: a short informal definition.
The purpose of a gloss is to state the **critical** and **distinguishing** characteristics that all instances of a concept have in common, in an informal but controlled way.

Gloss Formulation Guidelines:

Step 1: Start with the supertype of the concept being defined. e.g., "Object: An entity that…", "Physical Object: An object that …".

يبدأ التعريف بالجنس الأعلى للمفهوم المراد تعريفه.

entity شَيْء | كَيْنُونَة | كَائِن
أيّما وُجِد أو سيوجد ونستطيع إدراكه أو تخيّله
مثال: كلّ شيء على ما يرام
Whatever existed or will exist, and can be realized or imagined
293198

object | independent continuant مَوْجُود | كَائِن | قَائِم | حَقِيقِيّ | واقِعِيّ | شَيْء | ذَات | قَيُّومٌ
شيء له ذات مستقلّة بنفسه، وحاضر كليّاً في الزّمن، ويُدرك بذاته قياساً أو لذاته اعتباراً
مثال: يختلف إدراكنا لأيّ موجود لاختلاف ما يميّز أنواعه من الصّفات الجوهريّة
An entity that is wholly and independently present in time, and is realized either for its concrete or social existence
النوع : {شَيْء}
293200

physical object | material entity مَوْجُودٌ مَادِّيٌّ | مَحْسُوسٌ | مَلْمُوسٌ | كَائِنٌ | حَقِيقِيٌّ | مُجَسَّدٌ
موجود يشغل حيّزاً مكانيّاً، ويدرك بذاته بالحواس أو بأدوات القياس
مثال: لكلّ مَوْجُود مَادّيّ حجم يمكن قياسة أو حسابه مهما صغر أو كبر
An object that occupies space, and is realized by senses or measuring tools
النوع : {مَوْجُود} إجزاء
293254

organism كَائِنٌ حَيٌّ | مُتَعَضٍّ
موجود ماديّ قائم بذاته، يُظهر صفة الحياة، يقوم بمجموعة من الأنشطة والعمليات الحيوية، وهو قادر على التنفّس، النموّ، التكاثر، والتغذية
مثال: التكيّف هو قدرة الكائن الحيّ على الاستجابة للمؤثرات الطارئة بهدف البقاء
Physical object that can act independently constituted to carry on the activities of life by means of organs separate in function but mutually dependent; organisms have the ability to breathe, grow, feed, and reproduce.
النوع : {مَوْجُودٌ مَادّيٌّ}
293261

Anatomical Structure بناء حيوي | كيان حيوي | بناء تشريحي | تركيب حيوي | تركيب تشريحي
موجود مادي ثلاثي الإبعاد يدخل في تركيب الكائن الحي، مكوناته ناتجة عن الترجمة البنائية المنظّمة لجينوم الكائن الحي
مثال: نسخ الجينوم هي المرحلة الاولى لتكوين بناء حيوي مستنسخ

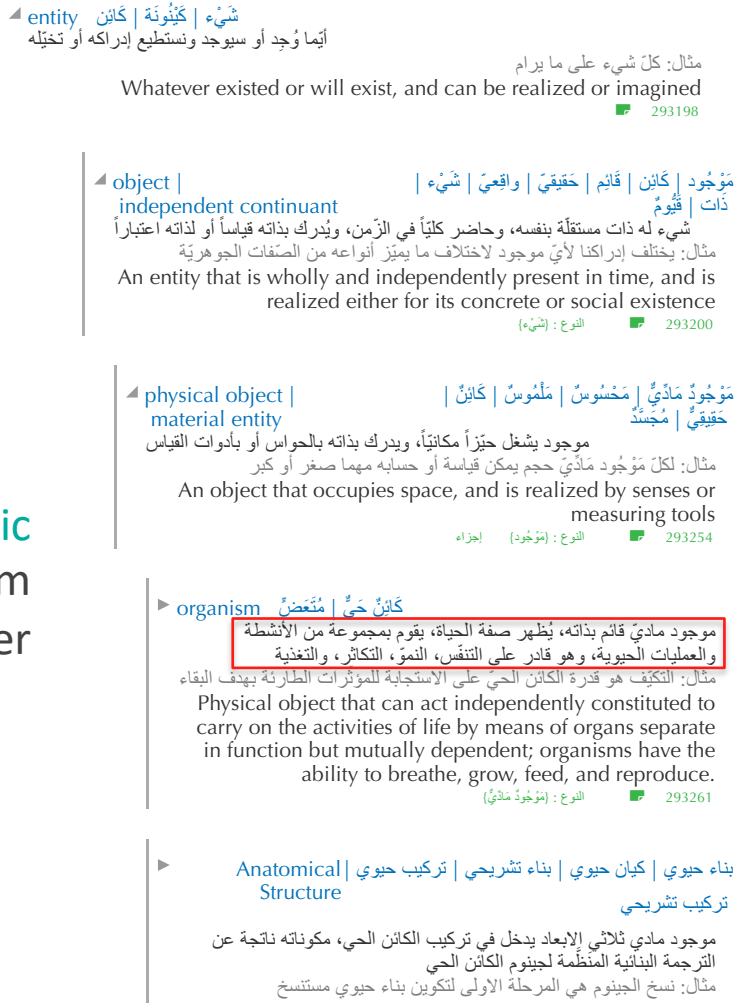
Gloss Formulation Methodology

Gloss: a short informal definition.
The purpose of a gloss is to state the **critical** and **distinguishing** characteristics that all instances of a concept have in common, in an informal but controlled way.

Gloss Formulation Guidelines:

Step 2: List only the most distinguishing and intrinsic characteristics that **specialize** the concept from its supertype, and that **differentiate** it from other concepts in the same level.

اذكر فقط **الخصائص الجوهرية المميزة** التي تخصص المفهوم عن نوعه الأعلى، وتميزه عن المفاهيم الأخرى في نفس المستوى.



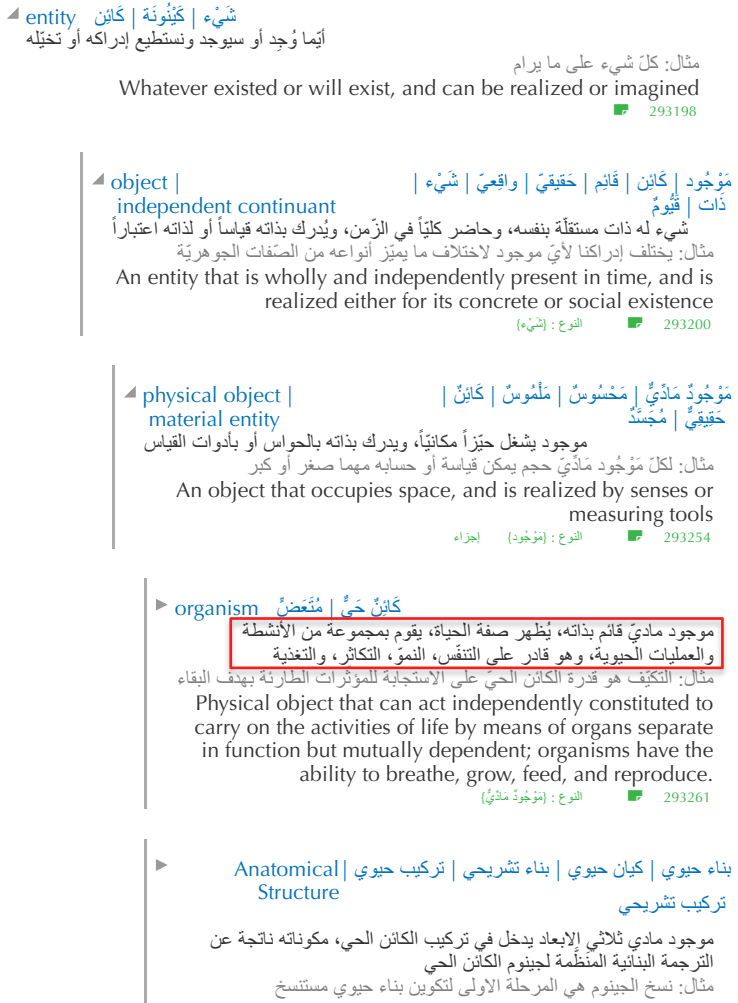
Gloss Formulation Methodology

Gloss: a short informal definition.
The purpose of a gloss is to state the **critical** and **distinguishing** characteristics that all instances of a concept have in common, in an informal but controlled way.

Gloss Formulation Guidelines:

Step 3: Write the distinguishing characteristics in the form of a sequence of propositions to help the reader to easily mentally rebuild the concept being defined in a declarative and non-narrative manner.

صياغة الخصائص المميزة على شكل **مجموعة من القضايا** لاستنباط المعنى، تسهيل إعادة بناء المفهوم ذهنيًا بطريقة تقريرية وغير سردية



Natural Language Processing

The Arabic Ontology

In this lecture:

- ❑ Part 1: Need for Linguistic Ontologies
- ❑ Part 2: General Overview
- ❑ Part 3: Upper-level Concepts
- ❑ Part 4: Gloss Formulation Guidelines
- ❑ Part 5: **Fundamentals and Formal Definitions**
- ❑ Part 6: Benchmarking Ontology Content
- ❑ Part 7: Discussion
- ❑ Part 8: Practice

Concept (vs Synset)

Definition: Concept (Jarrar, 2021):

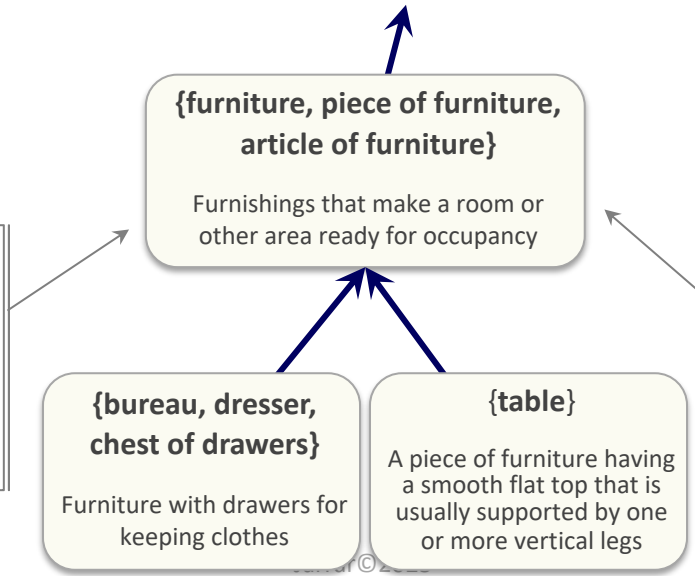
Given a concept c, its intensional interpretation c^I is defined on a *domain space* ⟨D, W⟩ as a function $c^I: W \rightarrow 2^D$, where D is a domain and W is a set of maximal states of affairs on D. For a concept c, the set $E_c = \{c^I(w) \mid w \in W\}$ is the set of the admissible extensions of c. Two concepts having the same set of *admissible* instances, in all states of affairs, are considered the same concept.

Wordnet

made of synsets
(linguistic concepts)

Synset: *signifies a concept*; a thought in our mind.

Individuals are also linguistic concepts (ISO 1087-1:2000)



Arabic Ontology

made of concepts
(classes of individuals).

Concept: *class of individuals*; characteristics its instances have in common.

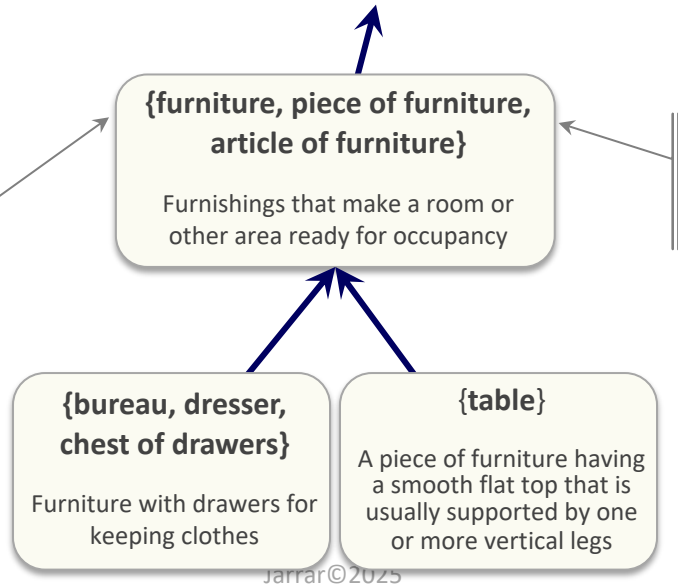
In BFO, Universal/Defined Class.

Synonymy Relation

Definition: Synonymy Relation (Jarrar, 2021)
Given two terms t_1 and t_2 lexicalizing concepts c_1 and c_2, respectively, then t_1 and t_2 are considered to be synonymous *iff* $c_1 = c_2$. In this way, synonymy can be defined as an equivalence relation $=_c$ between terms lexicalizing the same concept, thus it is a reflexive, symmetric and transitive relation.

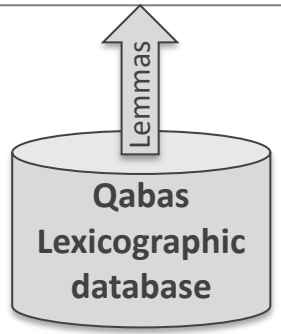
Wordnet

Synonyms: "two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value" (Miller et al., 1990).



Arabic Ontology

Synonyms: alternative labels/names of concepts.



Linking the semantic with the morphology levels

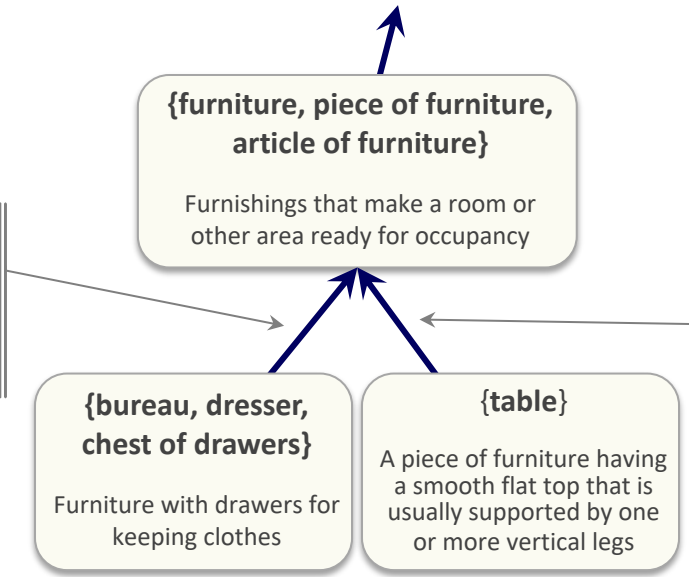
Subsumption Relation

Definition: Subsumption (Jarrar, 2021):

Concept c_1 subsumes concept c_2, *iff* every instance of c_2 is an instance of c_1, in every possible state of affairs.

Wordnet

Arabic Ontology



Hyponymy: If native speakers accept a sentence like: B is a kind of A

Subsumption: a subset relation between concepts. Every instance in c_1 is also an instance in c_2.

Parthood Relation

The parthood relationship is not fully formalized in the current version of the Arabic Ontology. Our progress in formalizing related relationships (e.g., **occurrent-part-of**, **temporalpart-of**, **isotypic-part-of**, **homeomeric-in**, **cumulative-with**, and **is-telic-in**) can be found in a previous work (Jarrar & Ceusters, 2017).

p **occurrent-part-of** *q*
a primitive relation of parthood holding independently of time between two process instances when one is a subprocess of the other (Arp et al 2015:135).

p **temporal-part-of** *q* =def.
p **occurrent-part-of** *q*
& for some temporal region *r* *p* spans *r*
& for all occurrents *c*, *r′*
if (*c* **spans** *r′* & *r′* **occurrent-part-of** *r*
then (*c* occurrent-part-of *p* iff *c* **occurrent-part-of** *q*))
(Smith 2012, corrected).

p **is-telic-in** *R* =def.
p instance-of *P*
& there exists some process *q* **instance-of** *Q* and some process *r* **instance-of** *R*, such that
(1) *q* not **instance-of** *P*,
(2) *p* not **instance-of** *Q*,
(3) *p* **precedes** *q*, and
(4) *p* and *q* are **temporal-part-of** *r*.

p **isotypic-part-of** *q* =def.
p **temporal-part-of** *q*
& *p* **instance-of** all types instantiated by *q*.

p **weakly-homeomeric-in** *P* =def.
all temporal parts of *p* which are not process boundaries are instances of *P*.

p **strongly-homeomeric-in** *P* =def.
all temporal parts of *p* which are not process boundaries are instances of *P* and there is no such part of *p* that instantiates a subtype of *P*

p **cumulative-with** *q* =def.
all process types instantiated by *p* and all process types instantiated by *q* are instantiated by *p*, *q* and *p*+*q*.

Natural Language Processing

The Arabic Ontology

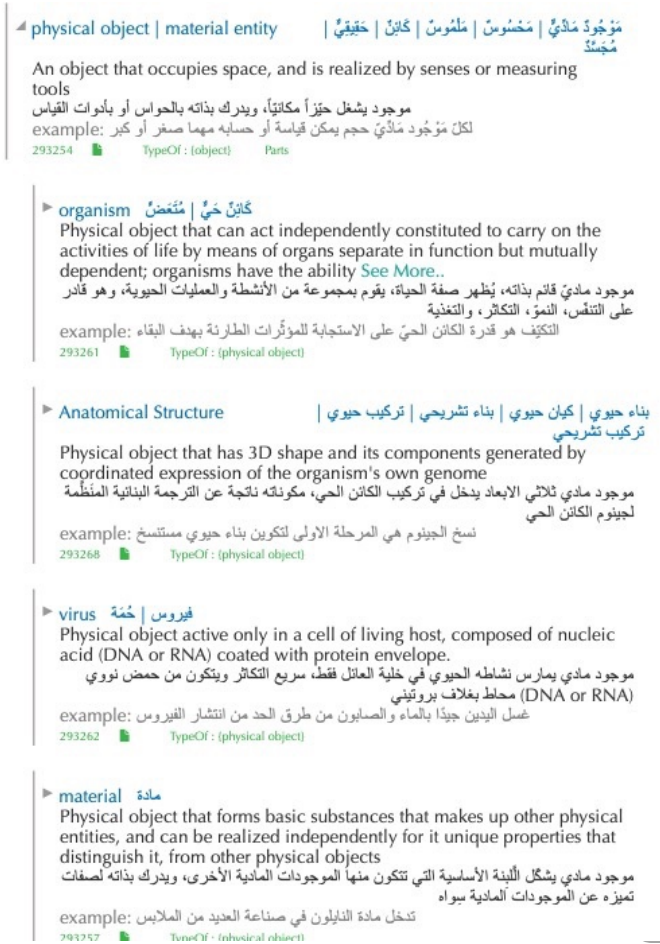
In this lecture:

- Part 1: Need for Linguistic Ontologies
- Part 2: General Overview
- Part 3: Upper-level Concepts
- Part 4: Gloss Formulation Guidelines
- Part 5: Fundamentals and Formal Definitions
- Part 6: **Benchmarking Ontology Content**
- Part 7: Discussion
- Part 8: Practice

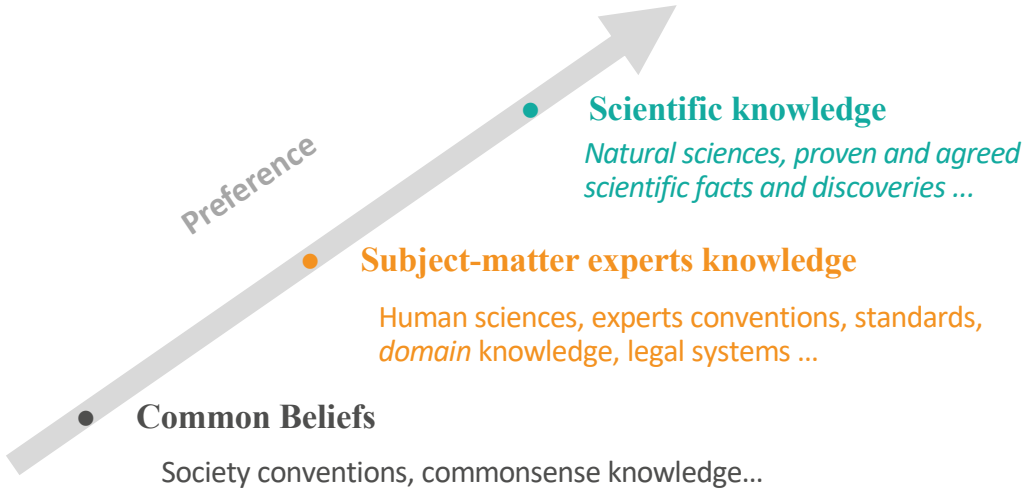
Benchmarking the ontology content

- **What should the ontology capture and adhere to?**
- **On what basis the correctness of the ontology content can be benchmarked?**

  - Should concepts be defined/classified based on what Arabic speakers commonly believe!
  - Should we adopt a certain lexicon and formalize it!
  - Should we rely on what the scientific literature accepts!
  - Should we build the ontology based on what we, the ontology builders, believe!



Benchmarking Methodology



The ontological precision of the classifications is benchmarked against the following, in order:

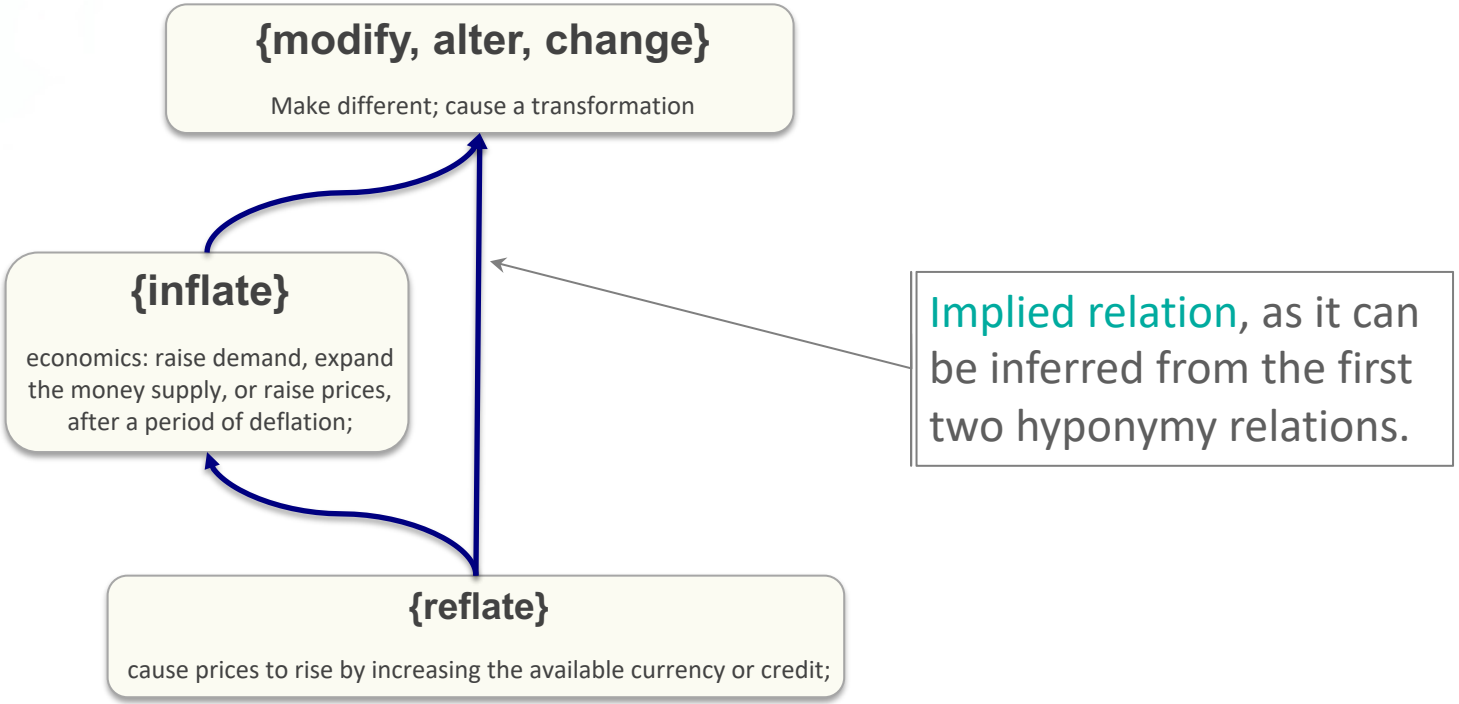
1. **Scientific knowledge**, which scientists typically accept on the basis of experimentation and verification and commonly agree about. If no mature answer is found in the state-of-art scientific discoveries, then against,
2. **Subject-matter experts' and domain knowledge and conventions**. If no answer can be synthesized or attained from experts' knowledge, then against,
3. **Commonsense knowledge**, repeatedly found in quality lexicons and literature.

More issues in wordnet:

Which might not be correct from a formal/ontological perspective

Formal/Ontological issues in wordnet

No benefits for including implied relations



Formal/Ontological issues in wordnet

Might be different linguistic concepts, but ontologically it is the same instance.

{evening star, Hesperus, Vesper}

a planet (usually Venus) seen at sunset in the western sky

{Phosphorus, Lucifer, daystar, morning star}

a planet (usually Venus) seen just before sunrise in the eastern sky

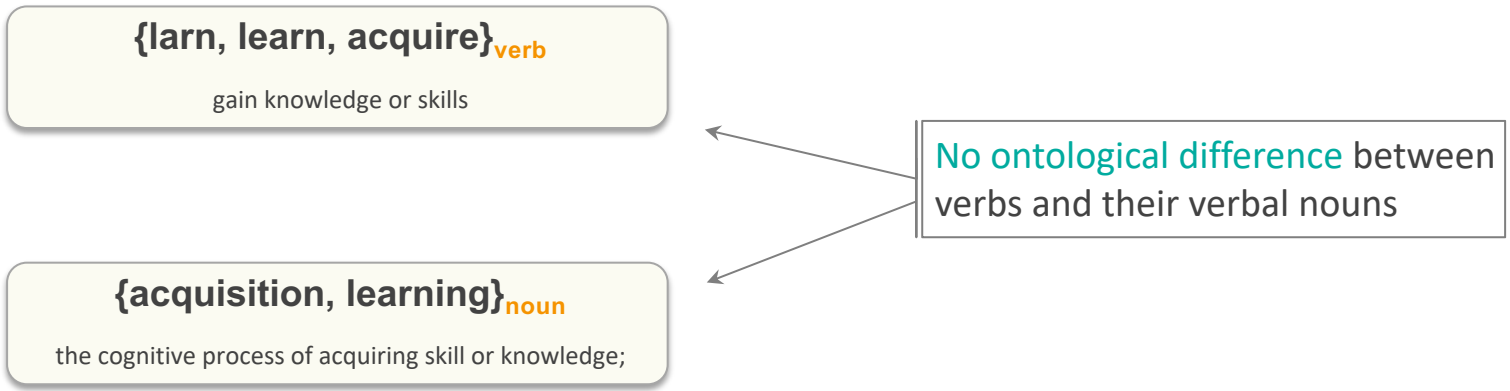
Same instance (i.e., Venus) that people see at different occasions.

Formal/Ontological issues in wordnet

Verbs are linguistic rather than ontological categories.

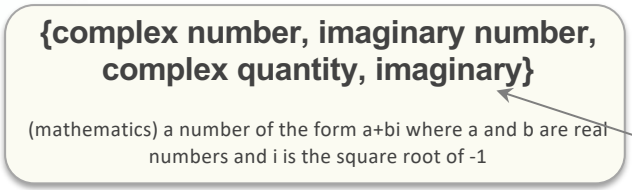
Ontologies capture the events that verbs denote rather than verbs themselves.

➔ We say (he learns, he learned, he is learning, the learning he ..,) referring to the exact same learning event.



Formal/Ontological issues in wordnet

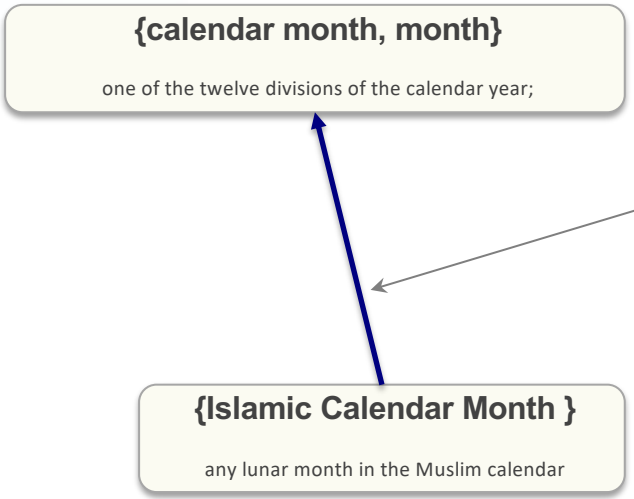
Who/How to decide on the accuracy?



not synonyms, Imaginary number is only a special case of a Complex number. Similarly, WordNet provides a poor classification of the types of numbers, e.g., Real, Rational, Natural, and Integer numbers are all subsumed by Number, while they subsume each other.

Formal/Ontological issues in wordnet

Who/How to decide on the accuracy?



Inaccurate, because Month is defined as one of the twelve divisions of the *calendar year*, which is a Gregorian year as defined in the gloss (i.e., an average month is about 30.43 days); however, an average of an Islamic month is 29.53 days. In fact, both months belong to two different calendar systems. A Gregorian month is 1/12 of a Gregorian year, while an Islamic year is a multiple of 12 lunar months.

Natural Language Processing

The Arabic Ontology

In this lecture:

- Part 1: Need for Linguistic Ontologies
- Part 2: General Overview
- Part 3: Upper-level Concepts
- Part 4: Gloss Formulation Guidelines
- Part 5: Fundamentals and Formal Definitions
- Part 6: Benchmarking Ontology Content
- Part 7: **Discussion**
- Part 8: Practice

Natural Language Processing

The Arabic Ontology

In this lecture:

- ❑ Part 1: Need for Linguistic Ontologies
- ❑ Part 2: General Overview
- ❑ Part 3: Upper-level Concepts
- ❑ Part 4: Gloss Formulation Guidelines
- ❑ Part 5: Fundamentals and Formal Definitions
- ❑ Part 6: Benchmarking Ontology Content
- ❑ Part 7: Discussion
- ❑ Part 8: **Practice**

References

1. Mustafa Jarrar: **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content.** Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021
2. Mustafa Jarrar, Tymaa Hammouda: **Qabas: An Open-Source Arabic Lexicographic Database.** In Proceedings of LREC-COLING 2024, pages 13363–13370, Torino, Italia. ELRA and ICCL.
3. Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammed Khalilia: **SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks.** Proceedings the 1st ArabicNLP, Part of the ACL 2023. ACL.
4. Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: **A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms.** In Proceedings of GWC2023, (pp.274-283). Spain, 2023
5. Sanad Malaysha, Mustafa Jarrar, Mohammed Khalilia: **Context-Gloss Augmentation for Improving Arabic Target Sense Verification.** In Proceedings of GWC2023, (pp.274-283). Spain, 2023
6. Moustafa Al-Hajj, Mustafa Jarrar: **ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD.** In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40--48, 2021
7. Moustafa Al-Hajj, Mustafa Jarrar: **LU-BZU at SemEval-2021 Task 2: Word2Vec and Lemma2Vec performance in Arabic Word-in-Context disambiguation.** In Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval2021) Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). PP 748--755, Association for Computational Linguistics. 2021
8. Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, Khaled Shaalan: **Extracting Synonyms from Bilingual Dictionaries.** The 11th International Global Wordnet Conference (GWC2021), Global Wordnet Association. (pp. 215-222). Pretoria, South Africa, 2021
9. Mustafa Jarrar, Hamzeh Amayreh: **An Arabic-Multilingual Database with a Lexicographic Search Engine.** The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Pages(234-246). LNCS 11608, Springer. 2019
10. Mustafa Jarrar, Hamzeh Amayreh, John P. McCrae: **Representing Arabic Lexicons in Lemon - a Preliminary Study.** The 2nd Conference on Language, Data and Knowledge (LDK 2019). Pages(29-33). CEUR, Volume 2402. ISSN:1613-0073. Leipzig, Germany. 2019
11. Diana Alhafi, Anton Deik, Mustafa Jarrar: **Usability Evaluation of Lexicographic e-Services.** The 16th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. Abu Dhabi, UAE. 2019
12. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: **Diacritic-Based Matching of Arabic Words.** ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1-10:21), ACM, ISSN:2375-4699. December, 2018
13. Mustafa Jarrar, Werner Ceusters: **Classifying Processes and Basic Formal Ontology.** Proceedings of the 8th International Conference on Biomedical Ontology (ICBO 2017), Newcastle, UK. 2017
14. Mustafa Jarrar: **Building a Formal Arabic Ontology (Invited Paper)**. In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso, Arab League. Tunis, July 26-28, 2011.
15. Mustafa Jarrar: **Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering.** In proceedings of the 15th International World Wide Web Conference (WWW2006). Edinburgh, Scotland. Pages 497-503. ACM Press. ISBN: 1595933239. May 2006.
16. Mustafa Jarrar, Anton Deik, Bilal Faraj: **Ontology-based Data and Process Governance Framework -The Case of e-Government Interoperability in Palestine**. Proceedings of the IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA'11). Pages(83-98). 2011.
17. Mustafa Jarrar and Robert Meersman: **Ontology Engineering -The DOGMA Approach**. Book Chapter in "Advances in Web Semantics I". Chapter 3. Pages 7-34. LNCS 4891, Springer. (2008).
18. Mustafa Jarrar: **Tutorial on Arabic Ontology Engineering**. The ACS/IEEE International Conference on Computer Systems and Applications. Tunis, 2017
19. Mustafa Jarrar, Maria Keet, and Paolo Dongilli: **Multilingual verbalization of ORM conceptual models and axiomatized ontologies** . Technical report. STARLab, Vrije Universiteit Brussel, February 2006.
20. Mustafa Jarrar: **Mapping ORM into the SHOIN/OWL Description Logic- Towards a Methodological and Expressive Graphical Notation for Ontology Engineering** . In OTM 2007 workshops: Proceedings of the International Workshop on Object-Role Modeling (ORM'07). Pages (729-741), LNCS 4805, Springer. ISBN: 9783540768890. Portogal. November, 2007
21. Tymaa Hammouda, Mustafa Jarrar, Mohammed Khalilia: **SinaTools: Open Source Toolkit for Arabic Natural Language Understanding.** In Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024), Procedia Computer Science, Dubai. ELSEVIER.
22. Mustafa Jarrar, Diyam Akra, Tymaa Hammouda: **ALMA: Fast Lemmatizer and POS Tagger for Arabic.** In Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024), Procedia Computer Science, Dubai. ELSEVIER.
23. Alaa Aljabari, Lina Duaibes, Mustafa Jarrar, Mohammed Khalilia: **Event-Arguments Extraction Corpus and Modeling using BERT for Arabic.** In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok, Thailand. Association for Computational Linguistics.
24. Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, Imed Zitouni: **ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task.** In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok, Thailand. Association for Computational Linguistics.
25. Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Muhammad Abdul-Mageed: **WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task.** In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok, Thailand. Association for Computational Linguistics.
26. Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi zaraket, Mohamad-Bassam Kurdy: **Nâbra: Syrian Arabic Dialects with Morphological Annotations.** In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the TKM24 2023. ACL.
27. Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem: **ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic.** In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the TKM24 2023. ACL.
28. Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood, Martin Waehlisch: **Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations.** The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. Egypt. DOI 10.1109/AICCSA59173.2023.10479250. 2023
29. Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, Fadi Zaraket: **Curras + Baladi: Towards a Levantine Corpus.** In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022