HOW FINE GRAINED CAN ENTITY TYPES GET?

Michael K. Bergman[1], Coralville, Iowa USA

March 8, 2016

AI3:::Adaptive Information blog

Entity recognition or extraction is a key task in natural language processing and one of the most common uses for knowledge bases. Entities are the unique, individual things in the world, and are also sometimes used to characterize some concepts [1]. Context plays an essential role in entity recognition. In general terms we may refer to a thing such as a *camera*; but a photographer may want more fine-grained distinctions such as *SLR cameras* or further sub-types like *digital SLR cameras* or even specific models like the *Canon EOS 7D Mark II* or even the name of the photographer's favorite camera, such as '*Shutter Sue*'. Capitalized names (as is the reference source for named entity recognition) often signals we are dealing with a true individual entity, but again, depending on context, a named automobile such as Chevy Malibu may refer to a specific car or to the entire class of Malibu cars.

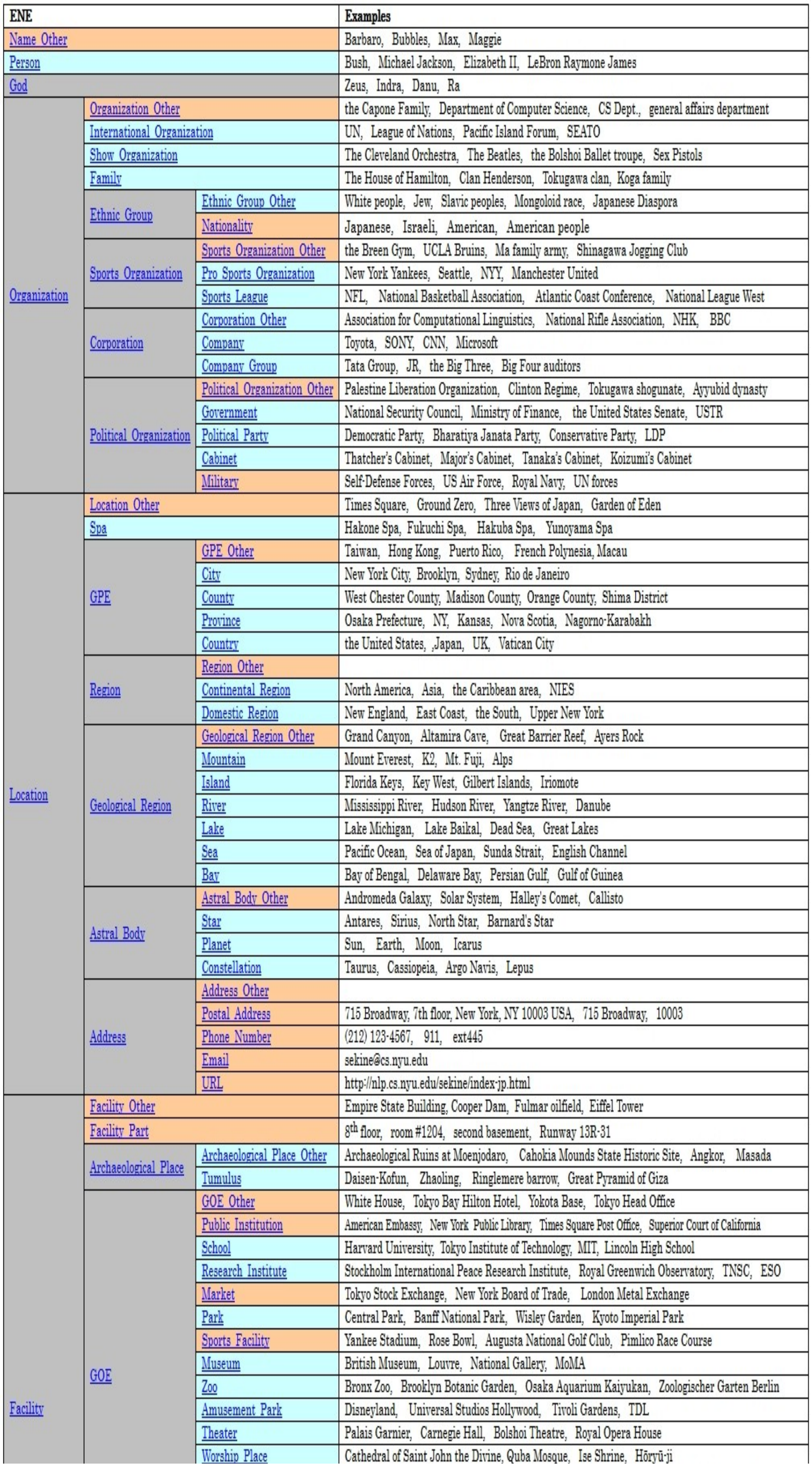
The "official" practice of named entity recognition began with the Message Understanding Conferences, especially MUC-6 and MUC-7, in 1995 and 1997. These conferences began competitions for finding "named entities" as well as the practice of in-line tagging [2]. Some of these accepted '*named entities*' are also written in lower case, with examples such as rocks ('gneiss') or common animals or plants ('daisy') or chemicals ('ozone') or minerals ('mica') or drugs ('aspirin') or foods ('sushi') or whatever. Some deference was given to the idea of Kripke's "rigid designators" as providing guidance for how to identify entities; rigid designators include proper names as well as certain natural kinds of terms like biological species and substances. Because of these blurrings, the nomenclature of "named entities" began to fade away. Some practitioners still use the term of **named entities**, though for some of the reasons outlined in this paper, Structured Dynamics prefers simply to use entity.

Much has changed in the twenty years since the seminal MUC conferences regarding entity recognition and characterization. We are learning to adopt a very fine-grained approach to entity types and a typology design suited to interoperating ("bridging") over a broad range of viewpoints and contexts. Most broadly, the idea of fine-grained entity types has led us to a logically grounded typology design.

The Growing Trend to Fine-Grained Entity Types

Beginning with the original MUC conferences, the initial entity types tested and recognized were for `person`, `organization`, and `location` names [3]. However, it did not take long for various groups and researchers to want more entity types, more distinctions. BBN categories, proposed in 2002, were used for question answering and consisted of 29 types and 64 subtypes [4]. Sekine put forward and refined over many years his Extended Entity Types, which grew to about 200 types [5], as shown in this figure:

[1] Email: mike@mkbergman.com

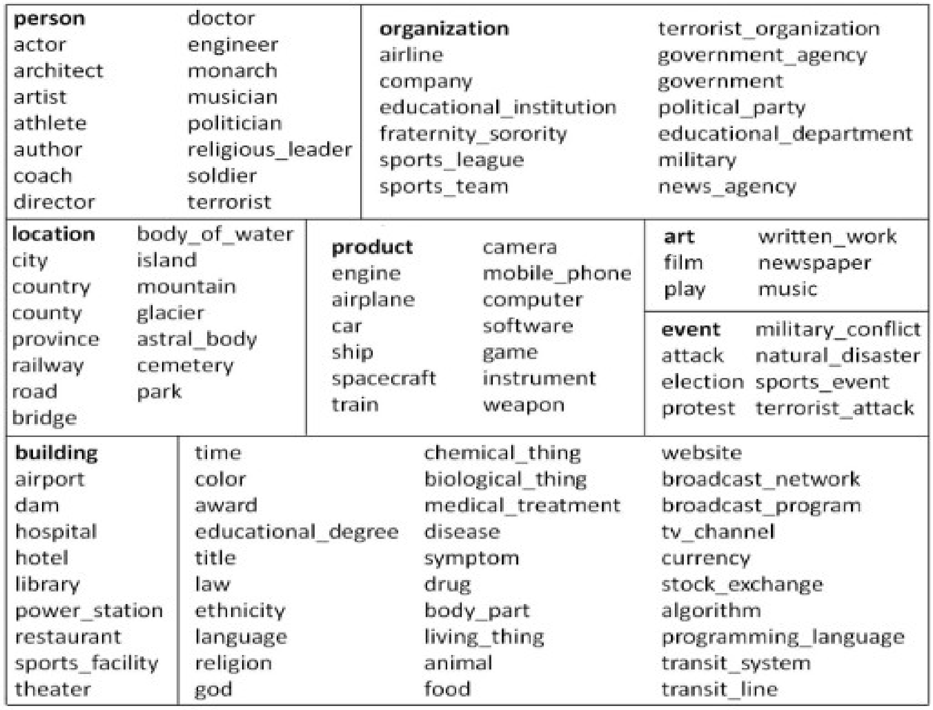


ENE			Examples
Name Other			Barbaro, Bubbles, Max, Maggie
Person			Bush, Michael Jackson, Elizabeth II, LeBron Raymone James
God			Zeus, Indra, Danu, Ra
Organization	Organization Other		the Capone Family, Department of Computer Science, CS Dept., general affairs department
	International Organization		UN, League of Nations, Pacific Island Forum, SEATO
	Show Organization		The Cleveland Orchestra, The Beatles, the Bolshoi Ballet troupe, Sex Pistols
	Family		The House of Hamilton, Clan Henderson, Tokugawa clan, Koga family
	Ethnic Group	Ethnic Group Other	White people, Jew, Slavic peoples, Mongoloid race, Japanese Diaspora
		Nationality	Japanese, Israeli, American, American people
	Sports Organization	Sports Organization Other	the Breen Gym, UCLA Bruins, Ma family army, Shinagawa Jogging Club
		Pro Sports Organization	New York Yankees, Seattle, NYY, Manchester United
		Sports League	NFL, National Basketball Association, Atlantic Coast Conference, National League West
	Corporation	Corporation Other	Association for Computational Linguistics, National Rifle Association, NHK, BBC
		Company	Toyota, SONY, CNN, Microsoft
		Company Group	Tata Group, JR, the Big Three, Big Four auditors
	Political Organization	Political Organization Other	Palestine Liberation Organization, Clinton Regime, Tokugawa shogunate, Ayyubid dynasty
		Government	National Security Council, Ministry of Finance, the United States Senate, USTR
		Political Party	Democratic Party, Bharatiya Janata Party, Conservative Party, LDP
		Cabinet	Thatcher's Cabinet, Major's Cabinet, Tanaka's Cabinet, Koizumi's Cabinet
		Military	Self-Defense Forces, US Air Force, Royal Navy, UN forces
Location	Location Other		Times Square, Ground Zero, Three Views of Japan, Garden of Eden
	Spa		Hakone Spa, Fukuchi Spa, Hakuba Spa, Yunoyama Spa
	GPE	GPE Other	Taiwan, Hong Kong, Puerto Rico, French Polynesia, Macau
		City	New York City, Brooklyn, Sydney, Rio de Janeiro
		County	West Chester County, Madison County, Orange County, Shima District
		Province	Osaka Prefecture, NY, Kansas, Nova Scotia, Nagorno-Karabakh
		Country	the United States, ,Japan, UK, Vatican City
	Region	Region Other	
		Continental Region	North America, Asia, the Caribbean area, NIES
		Domestic Region	New England, East Coast, the South, Upper New York
	Geological Region	Geological Region Other	Grand Canyon, Altamira Cave, Great Barrier Reef, Ayers Rock
		Mountain	Mount Everest, K2, Mt. Fuji, Alps
		Island	Florida Keys, Key West, Gilbert Islands, Iriomote
		River	Mississippi River, Hudson River, Yangtze River, Danube
		Lake	Lake Michigan, Lake Baikal, Dead Sea, Great Lakes
		Sea	Pacific Ocean, Sea of Japan, Sunda Strait, English Channel
		Bay	Bay of Bengal, Delaware Bay, Persian Gulf, Gulf of Guinea
	Astral Body	Astral Body Other	Andromeda Galaxy, Solar System, Halley's Comet, Callisto
		Star	Antares, Sirius, North Star, Barnard's Star
		Planet	Sun, Earth, Moon, Icarus
		Constellation	Taurus, Cassiopeia, Argo Navis, Lepus
	Address	Address Other	
		Postal Address	715 Broadway, 7th floor, New York, NY 10003 USA, 715 Broadway, 10003
		Phone Number	(212) 123-4567, 911, ext445
		Email	sekine@cs.nyu.edu
		URL	http://nlp.cs.nyu.edu/sekine/index-jp.html
Facility	Facility Other		Empire State Building, Cooper Dam, Fulmar oilfield, Eiffel Tower
	Facility Part		8th floor, room #1204, second basement, Runway 13R-31
	Archaeological Place	Archaeological Place Other	Archaeological Ruins at Moenjodaro, Cahokia Mounds State Historic Site, Angkor, Masada
		Tumulus	Daisen-Kofun, Zhaoling, Ringlemere barrow, Great Pyramid of Giza
	GOE	GOE Other	White House, Tokyo Bay Hilton Hotel, Yokota Base, Tokyo Head Office
		Public Institution	American Embassy, New York Public Library, Times Square Post Office, Superior Court of California
		School	Harvard University, Tokyo Institute of Technology, MIT, Lincoln High School
		Research Institute	Stockholm International Peace Research Institute, Royal Greenwich Observatory, TNSC, ESO
		Market	Tokyo Stock Exchange, New York Board of Trade, London Metal Exchange
		Park	Central Park, Banff National Park, Wisley Garden, Kyoto Imperial Park
		Sports Facility	Yankee Stadium, Rose Bowl, Augusta National Golf Club, Pimlico Race Course
		Museum	British Museum, Louvre, National Gallery, MoMA
		Zoo	Bronx Zoo, Brooklyn Botanic Garden, Osaka Aquarium Kaiyukan, Zoologischer Garten Berlin
		Amusement Park	Disneyland, Universal Studios Hollywood, Tivoli Gardens, TDL
		Theater	Palais Garnier, Carnegie Hall, Bolshoi Theatre, Royal Opera House
		Worship Place	Cathedral of Saint John the Divine, Quba Mosque, Ise Shrine, Hōryū-ji

Sekine Extended Entity Types

These ideas of extended entity types helped inform a variety of tagging services over the past decade, notably including OpenCalais, Zemanta, AlchemyAPI, and OpenAmplify, among others. Moreover the research community also expanded its efforts into more and more entity types, or what came to be known as fine-grained entities [6].

Some of these produced more formal organizations of entity type classifications. This one, from Ling and Weld proposed 112 entity types in 2012 [7]:



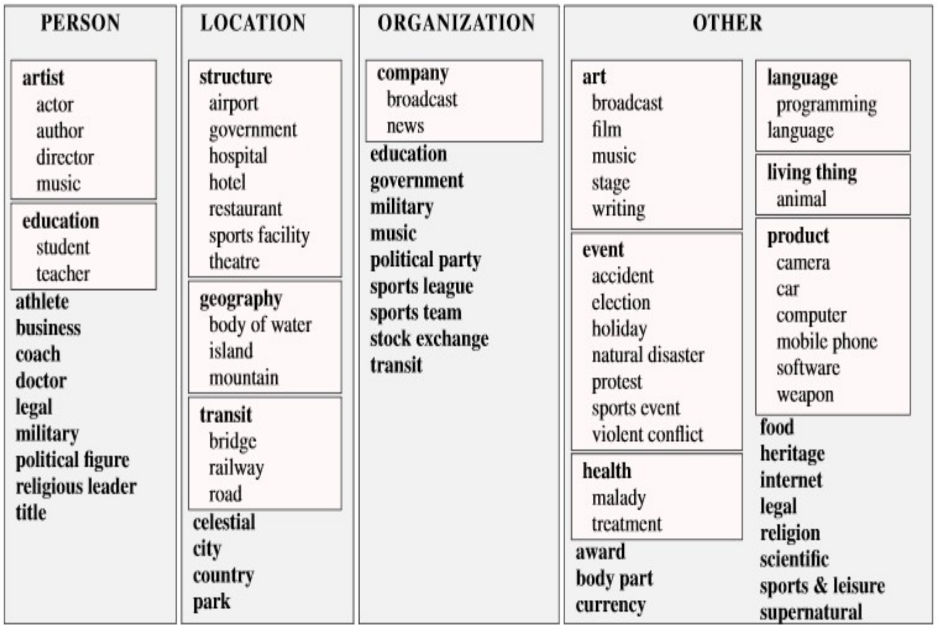
person	doctor	**organization**	terrorist_organization
actor	engineer	airline	government_agency
architect	monarch	company	government
artist	musician	educational_institution	political_party
athlete	politician	fraternity_sorority	educational_department
author	religious_leader	sports_league	military
coach	soldier	sports_team	news_agency
director	terrorist		

location	body_of_water	**product**	camera	**art**	written_work
city	island	engine	mobile_phone	film	newspaper
country	mountain	airplane	computer	play	music
county	glacier	car	software	**event**	military_conflict
province	astral_body	ship	game	attack	natural_disaster
railway	cemetery	spacecraft	instrument	election	sports_event
road	park	train	weapon	protest	terrorist_attack
bridge					

building	time	chemical_thing	website
airport	color	biological_thing	broadcast_network
dam	award	medical_treatment	broadcast_program
hospital	educational_degree	disease	tv_channel
hotel	title	symptom	currency
library	law	drug	stock_exchange
power_station	ethnicity	body_part	algorithm
restaurant	language	living_thing	programming_language
sports_facility	religion	animal	transit_system
theater	god	food	transit_line

Ling 112 Entity Types

Another one, from Gillick *et al.* in 2014 proposed 86 entity types [8], organized, in part, according to the same `person`, `organization`, and `location` types from the earliest MUC conferences:



PERSON
- artist
  - actor
  - author
  - director
  - music
- education
  - student
  - teacher
- athlete
- business
- coach
- doctor
- legal
- military
- political figure
- religious leader
- title

LOCATION
- structure
  - airport
  - government
  - hospital
  - hotel
  - restaurant
  - sports facility
  - theatre
- geography
  - body of water
  - island
  - mountain
- transit
  - bridge
  - railway
  - road
- celestial
- city
- country
- park

ORGANIZATION
- company
  - broadcast
  - news
- education
- government
- military
- music
- political party
- sports league
- sports team
- stock exchange
- transit

OTHER
- art
  - broadcast
  - film
  - music
  - stage
  - writing
- event
  - accident
  - election
  - holiday
  - natural disaster
  - protest
  - sports event
  - violent conflict
- health
  - malady
  - treatment
- award
- body part
- currency
- language
  - programming language
- living thing
  - animal
- product
  - camera
  - car
  - computer
  - mobile phone
  - software
  - weapon
- food
- heritage
- internet
- legal
- religion
- scientific
- sports & leisure
- supernatural

Gillick 86 Entity Types

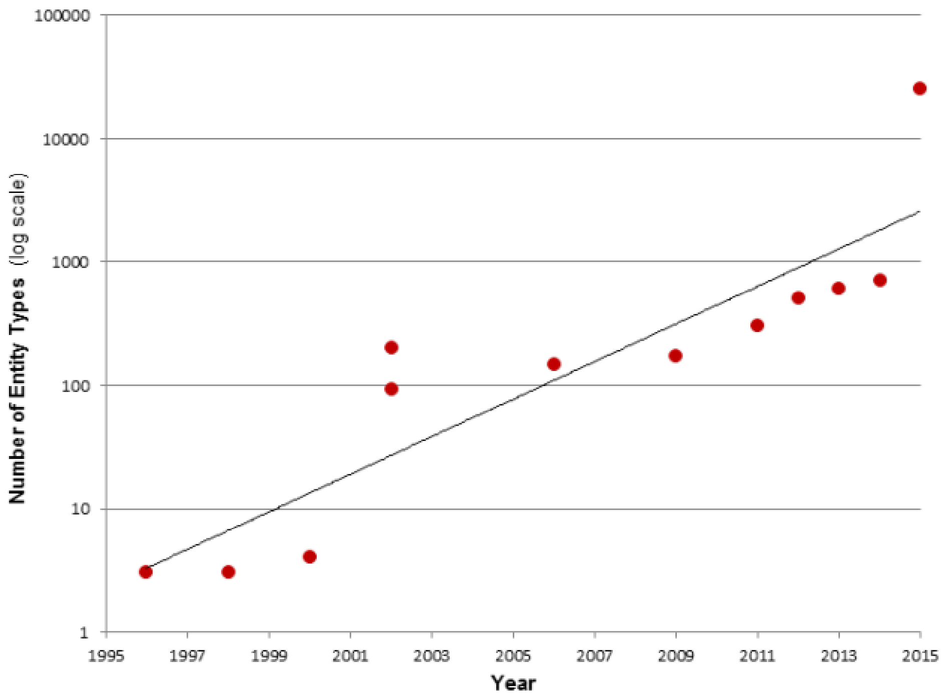
These efforts are also notable because machine learners have been trained to recognize the types shown. What entity types are covered, the different conceptions of the world, and how to organize entity types varies broadly across these references.

The complement to entity extraction for unstructured text is to label the text in the first place. For this, a number of schema presently exist that provide vocabularies of entity types and standard means for tagging text. These include:

- DBpedia Ontology: 738 types [9]
- schema.org: 636 types [10]
- YAGO: 505 types; see also HYENA [11]
- GeoNames: 654 "feature codes" [12]

In Structured Dynamics' own work, we have mapped the UMBEL knowledge graph against Wikipedia content and found that 25,000 nodes, or more than 70 percent of its 35,000 reference concepts, correspond to entity types [13]. These mappings provide typing connections for millions of Wikipedia articles. The typing and organization of entity types thus appears to be of enormous importance in modeling and leveraging the use of knowledge bases.

When we track the coverage of entity types over the past two decades we see logarithmic growth [13]:



Growth in Recognition of Entity Types

This growth in entity types comes from wanting to describe and organize things with more precision. Tagging and extracting structured information from text are obviously a key driver. Yet, for a given enterprise, what is of interest — and at what depth — for a particular task varies widely.

The fact that knowledge bases, such as Wikipedia (but, the lesson applies to domain-specific ones as well), can be supported by entity-level information for literally thousands of entity types means that rich information is available for driving the finest of fine-grained entity extractors. To leverage this raw, informational horsepower it is essential to have a grounded understanding of what an entity is, how to organize them into logical types, and an intensional understanding of the attributes and characteristics that allow inferencing to be conducted over these types. These understandings, in turn, point to the features that are useful to machine learners for artificial intelligence. These understandings also can inform a flexible design for accommodating entity types from coarse- to fine-grained, with variable depth depending on the domain of interest.

Natural Classes and Typologies

We take a realistic view of the world. That is, we believe that what we perceive in the world is real — it is not just a consequence of what we perceive and can be aware of in our minds [14] — and that there are forces and relationships in the world independent of us as selves. Realism is a longstanding tradition in philosophy that extends back to Aristotle and embraces, for example, the natural classification systems of living things as espoused by taxonomists such as Agassiz and Linnaeus.

Charles Sanders Peirce, an American logician and scientist of the late 19th and early 20th centuries, embraced this realistic philosophy but also embedded it in a belief that our understanding of the world is fallible and that we needed to test our perceptions via logic (the scientific method) and shared consensus within the community. His overall approach is known as pragmatism and is firmly grounded in his views of logic and his theory of signs (called semiotics or semeiotics). While there is absolute truth, it actually acts more as a limit, to which our seeking of additional knowledge and clarity of communication with language continuously approximates. Through the scientific method and questioning we get closer and closer to the truth and to an ability to communicate it to one another. But new knowledge may change those understandings, which in any case will always remain proximate.

Peirce's own words can better illustrate his perspective [15], some of which I have discussed elsewhere under his idea of "natural classes" [16]:

> "Thought is not necessarily connected with a brain. It appears in the work of bees, of crystals, and throughout the purely physical world; and one can no more deny that it is really there, than that the colors, the shapes, etc., of objects are really there." (Peirce CP 4.551)
> "What if we try taking the term "natural," or "real, class" to mean a class of which all the members owe their existence as members of the class to a common final cause? This is somewhat vague; but it is better to allow a term like this to remain vague, until we see our way to rational precision." (Peirce CP 1.204)
> ". . . it may be quite impossible to draw a sharp line of demarcation between two classes, although they are real and natural classes in strictest truth. Namely, this will happen when the form about which the individuals of one class cluster is not so unlike the form about which individuals of another class cluster but that variations from each middling form may precisely agree." (Peirce CP 1.208)
> "When one can lay one's finger upon the purpose to which a class of things owes its origin, then indeed abstract definition may formulate that purpose. But when one cannot do that, but one can trace the genesis of a class and ascertain how several have been derived by different lines of descent from one less specialized form, this is the best route toward an understanding of what the natural classes are." (Peirce CP 1.208)
> "The descriptive definition of a natural class, according to what I have been saying, is not the essence of it. It is only an enumeration of tests by which the class may be recognized in any one of its members. A description of a natural class must be founded upon samples of it or typical examples." (Peirce CP 1.223)
> "Natural classes" thus are a testable means to organize the real objects in the world, the individual particulars of what we call "entities". In Structured Dynamics' usage, we define an ***entity*** as something that is an individual object, either real or mental such as an idea, either a part or a whole, and that has:
>
> - identity, which can be referred to via symbolic names
> - context in relation to other objects, and
> - characteristic attributes, with some expressing the essence of what type of object it is.

The key to classification of entities into categories (or "types" as we use herein) is based on this intensional understanding of attributes. Further, Peirce was expansive in his recognition of what kinds of objects could be classified, specifically including ideas, with application to areas such as social classes, man-made objects, the sciences, chemical elements and living organisms [17]. Again, here are some of Peirce's own words on the classification of entities [15]:

> "All classification, whether artificial or natural, is the arrangement of objects according to ideas. A natural classification is the arrangement of them according to those ideas from which their existence results." (Peirce CP 1.231)
> "The natural classification of science must be based on the study of the history of science; and it is upon this same foundation that the alcove-classification of a library must be based." (Peirce CP 1.268)
> "All natural classification is then essentially, we may almost say, an attempt to find out the true genesis of the objects classified. But by genesis must be understood, not the efficient action which produces the whole by producing the parts, but the final action which produces the parts because they are needed to make the whole. Genesis is production from ideas. It may be difficult to understand how this is true in the biological world, though there is proof enough that it is so. But in regard to science

> it is a proposition easily enough intelligible. A science is defined by its problem; and its problem is clearly formulated on the basis of abstracter science." (Peirce CP 1.227)
> A natural classification system is one, then, that logically organizes entities with shared attributes into a hierarchy of types, with each type inheriting attributes from its parents and being distinguished by what Peirce calls its *final cause*, or purpose. This hierarchy of types is thus naturally termed a ***typology***.

An individual that is a member of a natural class has the same kinds of attributes as other members, all of which share this essence of the final cause or purpose. We look to Peirce for the guidance in this area because his method of classification is testable, based on discernable attributes, and grounded in logic. Further, that logic is itself grounded in his theory of signs, which ties these understandings ultimately to natural language.

Logic and the Typology Design

Unlike more interconnected knowledge graphs (which can have many network linkages), typologies are organized strictly along these lines of shared attributes, which is both simpler and provides an orthogonal means for investigating type class membership. Further, because the essential attributes or characteristics across entities in an entire domain can differ broadly — such as living v inanimate things, natural things v man-made things, ideas v physical objects, etc. — it is possible to make disjointedness assertions between entire groupings of natural entity classes. Disjoint assertions combined with logical organization and inference mean a typology design that lends itself to reasoning and tractability.

The idea of nested, hierarchical types organized into broad branches of different entity typologies also provides a very flexible design for interoperating with a diversity of world views and degrees of specificity. The photographer, as I discussed above, is interested in different camera types and even how specific cameras can relate to a detailed entity typing structure. Another party more interested in products across the board may have a view to greater breadth, but lesser depth, about cameras and related equipment. A typology design, logically organized and placed into a consistent grounding of attributes, can readily interoperate with these different world views.

A typology design for organizing entities can thus be visualized as a kind of accordion or squeezebox, expandable when detail requires, or collapsed to more coarse-grained when relating to broader views. The organization of entity types also has a different structure than the more graph-like organization of higher-level conceptual schema, or knowledge graphs. In the cases of broad knowledge bases, such as UMBEL or Wikipedia, where 70 percent or more of the overall schema is related to entity types, more attention can now be devoted to aspects of concepts or relations.

The idea that knowledge bases can be purposefully crafted to support knowledge-based artificial intelligence, or KBAI, flows from these kinds of realizations. We begin to see that we can tease out different aspects of a knowledge base, each with its own logic and relation to the other aspects. Concepts, entities, attributes and relations — including the natural classes or types that can logically organize them — all deserve discrete attention and treatment.

Peirce's consistent belief that the real world can be logically conceived and organized provides guidance for how we can continue to structure our knowledge bases into computable form. We now have a coherent base for treating entities and their natural classes as an essential component to that thinking. We can continue to be more fine-grained so long as there are unique essences to things that enable them to be grouped into natural classes.

Acknowledgements

This article was originally posted on the *AI3::Adaptive Information* Web site at http://www.mkbergman.com/1927/how-fine-grained-can-entity-types-get/. This version has been edited and

reformatted slightly for PDF distribution. We thank Cognonto Corporation for making this content freely available.

---

[1] The role for the label "entity" can also refer to what is known as the root node in some systems such as SUMO (see also http://virtual.cvut.cz/kifb/en/toc/229.html). In the OWL language and RDF data model we use, the root node is known as "thing". Clearly, our use of the term "entity" is much different than SUMO and resides at a subsidiary place in the overall TBox hierarchy. In this case, and frankly for most semantic matches, equivalences should be judged with care, with context the crucial deciding factor.

[2] N. Chinchor, 1997. "Overview of MUC-7," *MUC-7 Proceedings*, 1997.

[3] While all of these are indeed entity types, the early MUCs also tested dates, times, percentages, and monetary amounts.

[4] Ada Brunstein, 2002. "Annotation Guidelines for Answer Types". *LDC Catalog*, Linguistic Data Consortium. Aug 3, 2002.

[5] See the Sekine Extended Entity Types; the listing also includes attributes info at bottom of source page.

[6] For example, try this query, https://scholar.google.com/scholar?q="fine-grained+entity", also without quotes.

[7] Xiao Ling and Daniel S. Weld, 2012. "Fine-Grained Entity Recognition," in *AAAI*. 2012.

[8] Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh, 2104. "Context-Dependent Fine-Grained Entity Type Tagging," *arXiv preprint arXiv*:1412.1820 (2014).

[9] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann, 2009. "DBpedia-A Crystallization Point for the Web of Data." *Web Semantics: science, services and agents on the world wide web* 7, no. 3 (2009): 154-165; 170 classes in this paper. That has grown to more than 700; see http://mappings.dbpedia.org/server/ontology/classes/ and http://wiki.dbpedia.org/services-resources/datasets/dataset-2015-04/dataset-2015-04-statistics.

[10] The listing is under some dynamic growth. This is the official count as of September 8, 2015, from http://schema.org/docs/full.html. Current updates are available from Github.

[11] Joanna Biega, Erdal Kuzey, and Fabian M. Suchanek, 2013. "Inside YAGO2: A Transparent Information Extraction Architecture," in *Proceedings of the 22nd international conference on World Wide Web*, pp. 325-328. International World Wide Web Conferences Steering Committee, 2013. Also see Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, Gerhard Weikum, 2012. "HYENA: Hierarchical Type Classification for Entity Names," in *Proceedings of the 24th International Conference on Computational Linguistics, Coling 2012*, Mumbai, India, 2012.

[12] See https://en.wikipedia.org/wiki/GeoNames.

[13] This figure and some of the accompanying text comes from a prior article, M.K. Bergman, "Creating a Platform for Machine-based Artificial Intelligence", *AI3:::Adaptive Information* blog, September 21, 2015.

[14] Realism is often contrasted to idealism, nominalism or conceptualism, wherein how the world exists is a function of how we think about or name things. Descartes, for example, summarized his conceptualist view with his aphorism "I think, therefore I am."

[15] See the electronic edition of *The Collected Papers of Charles Sanders Peirce*, reproducing *Vols. I-VI*, Charles Hartshorne and Paul Weiss, eds., 1931-1935, Harvard University Press, Cambridge, Mass., and Arthur W. Burks, ed., 1958, *Vols. VII-VIII*, Harvard University Press, Cambridge, Mass. The citation scheme is volume number using Arabic numerals followed by section number from the collected papers, shown as, for example, CP 1.208.

[16] M.K. Bergman, 2015. "'Natural' Classes in the Knowledge Web", *AI3:::Adaptive Information* blog, July 13, 2015.

[17] See, for example, Menno Hulswit, 2000. "Natural Classes and Causation", in the online *Digital Encyclopedia of Charles S. Peirce*.