Available Article

Author's final: This draft is prior to submission for publication, and the subsequent edits in the published version. If quoting or citing, please refer to the proper citation of the published version below to check accuracy and pagination.

Cite as: Bergman, M. K. Appendix B: The KBpedia Resource. in *A Knowledge Representation Practionary: Guidelines Based on Charles Sanders Peirce* (ed. Bergman, M. K.) 409-419 (Springer International Publishing, 2018). doi:10.1007/978-3-319-98092-8_19

Official site: https://link.springer.com/book/10.1007/978-3-319-98092-8

Full-text: http://www.mkbergman.com/publications/akrp/appendix-b.pdf

Abstract: KBpedia is structured to enable useful splits across a myriad of dimensions from entities to relations to types that can all be selected to create positive and negative training sets, across multiple perspectives. The coherency of KBpedia provides a basis for logic tests to further improve accuracy, including the creation of local gold standards at an acceptable cost. KKO sets the umbrella structure for how KBpedia's six constituent knowledge bases are related to the system. We split KBpedia knowledge graph into concepts and topics, entities, events, attributes, annotations, and relations and their associated natural classifications or types.

APPENDIX B:

THE KBPEDIA RESOURCE

KBpedia is a computable knowledge structure resulting from the combined mapping of six, large-scale, public knowledge bases — Wikipedia, Wikidata, OpenCyc, GeoNames, DBpedia, and UMBEL.[1] The KBpedia structure separately captures entities, attributes, relations, and concepts. KBpedia classes these into a natural and rich diversity of types, with their meaning and relationships, logically and coherently organized into about 80 typologies.

KBpedia is the first full-blown ontology based on Charles Sanders Peirce's universal categories and logic of relations. The KBpedia knowledge structure is written in the OWL semantic language; all underlying structures are represented in either OWL or RDF. KBpedia follows best practices, many of which were pioneered by KBpedia's editors, governing knowledge and concept representation and annotations. All languages and knowledge representations are written in W3C-compliant standards.

The focal objective of KBpedia is to exploit large, public knowledge bases to support artificial intelligence using both supervised and unsupervised machine learning methods. KBpedia is explicitly designed to expose rich and meaningful feature sets to support the broadest range of machine learning methods. KBpedia is also specifically structured to enable useful splits across a myriad of dimensions from entities to relations to types that can all be selected to create positive and negative training sets, across multiple perspectives. The disjointedness of the SuperTypes that organize the 55,000 entity types in KBpedia provides a powerful selection and testing mechanism.[2] The coherency of KBpedia provides a basis for logic tests to further improve accuracy, including the creation of local gold standards, at an acceptable cost.

KBpedia is a continually evolving reference structure for knowledge representation and management. KBpedia is staged to provide working levels of interoperability for the linked data ecosystem. Artificial intelligence (AI) and machine learning are revolutionizing knowledge systems. The most important factor in knowledge-based AI's renaissance has been the availability of massive digital datasets for the training of machine learners. Wikipedia and data from search engines are central to

recent breakthroughs. Wikipedia is at the heart of Siri, Cortana, the former Freebase, DBpedia, Google's Knowledge Graph and IBM's Watson, to name just a prominent few AI question answering or virtual assistant systems. Natural language understanding is showing impressive gains across a range of applications. To date, all of these examples have been the result of bespoke efforts leveraging Wikipedia, in whole or part. The tens of millions of instances captured by Wikidata adds an entire ABox component to the knowledge structure. It is costly for standard enterprises to leverage these knowledge resources on their own.

Today's practices for leveraging these resources pose significant upfront and testing effort. Much latent knowledge remains unexpressed and not readily available to learners; it must be exposed, cleaned and vetted. We need to spend further upfront effort on selecting the features (variables) used and then to label the positive and negative training sets accurately. Without 'gold standards' — at still more cost — it is difficult to tune and refine the learners. The cost to develop tailored extractors, taggers, categorizers, and natural language processors is too high. KBpedia is meant to systematize a starter reference structure that new users may tailor to local domains at lower costs. Users may then apply integration and interoperability to structured, semi-structured and unstructured data; that is, everything from text to databases. KBpedia proves that existing knowledge bases can be staged to automate much of the tedium and reduce the costs now required to set up and train machine learners for knowledge purposes. Besides labeled training sets for supervised machine learning, KBpedia, with its rich feature sets across all aspects of the knowledge structure, is also an excellent basis for selecting training corpora for unsupervised learning. It is often advisable to include some initial unsupervised learning in a more general supervised learning context.

COMPONENTS

KBpedia is organized into a knowledge graph, KKO, the KBpedia Knowledge Ontology, with an upper structure based on Peircean logic. KKO sets the umbrella structure for how KBpedia's six constituent knowledge bases are related to the system. One of the three major branches of KKO, the Generals, represents the types in the system, with about 85% of the KBpedia's reference concepts residing there. These RCs are themselves entity types — that is, 47,000 natural classes of similar entities such as 'astronauts' or 'breakfast cereals' — which we organize into about 30 'core' typologies (among the 80 or so) that are mostly disjoint (non-overlapping) with one another. The typologies provide a flexible means for slicing-and-dicing the knowledge structure; the entity types provide the tie-in points to KBpedia's millions of individual instances. The separate *Glossary* defines many of the terms used by KBpedia; *Chapter 8* provides a more detailed discussion of KBpedia's vocabulary.

The KBpedia Knowledge Ontology (KKO)

We inform the upper structure that is the KBpedia Knowledge Ontology (KKO) using the triadic logic and universal categories of Charles Sanders Peirce. This trichotomy, also the basis for his views on semiosis (or the nature of signs), was in Peirce's view the most primitive or reduced manner by which to understand and categorize things, concepts, and ideas. We devote *Chapter 6* to the universal categories and touch upon them and semiosis throughout the main book. We express KBpedia's knowledge base grammar in the semantic Web language of OWL 2. Thus, we may apply most W3C standards to the KBpedia structure. The resulting, combined structure, as shown in *Figure B-1*, brings consistency across all source knowledge bases.

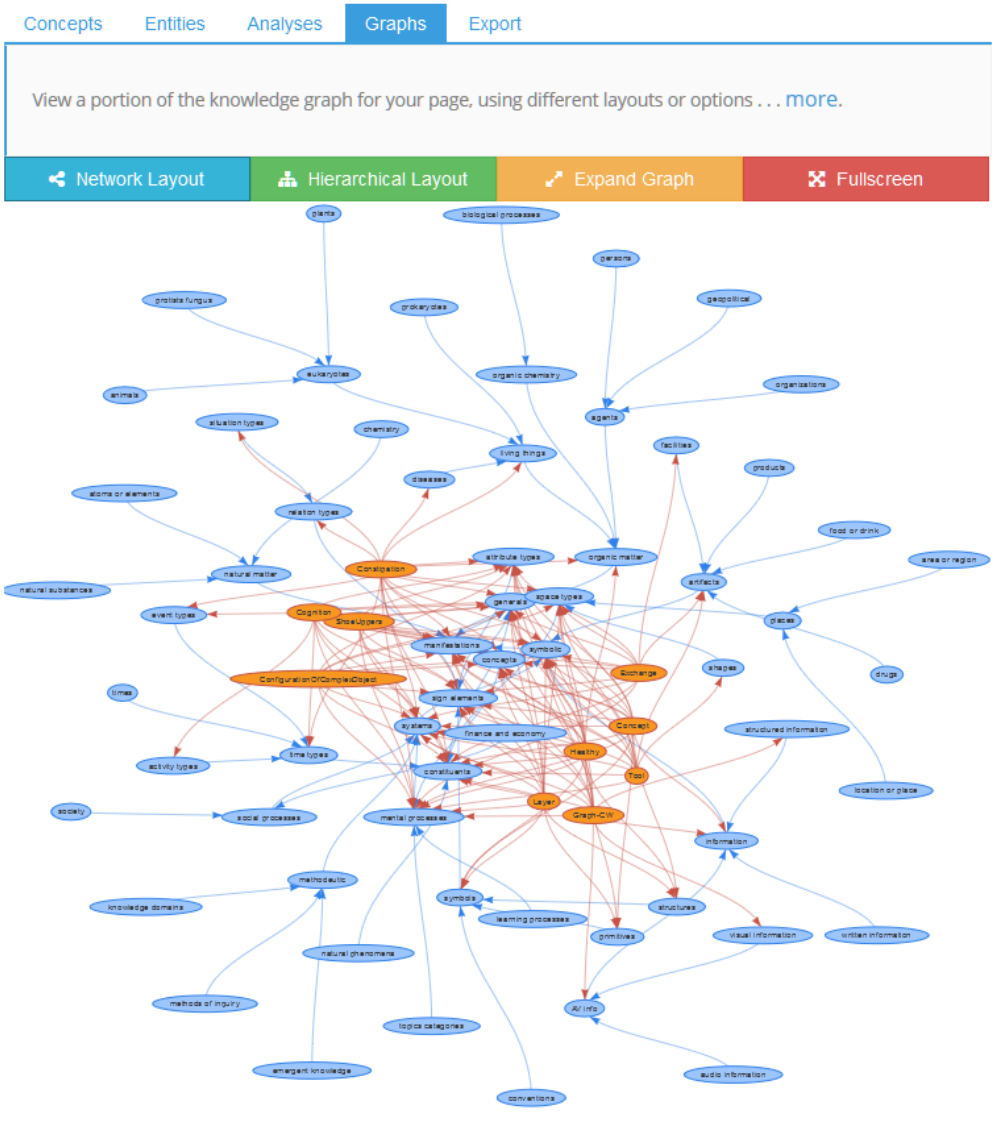

Figure B-1: KBpedia (KKO) Network Graph

This diagram, drawn from KBpedia's online demo, shows the main topic areas, or typologies, that tie into KKO, which we list in their entirety in the *Structure* section below. The structure of KKO makes it a computable knowledge graph that supports inference, reasoning, aggregations, restrictions, intersections, and other logical operations. KKO's logic basis provides a powerful way to represent individual things, classes of things, and how those things may combine or emerge as new knowledge.[3]

The KBpedia Knowledge Bases

Six, large-scale public knowledge bases are central to the KBpedia knowledge structure. These six sources are:

- Wikipedia - five million articles that capture the fundamental concepts and entities of basic human knowledge, often including structured data and with many linkages;
- Wikidata - structured data records for about 40 million individual entities;
- OpenCyc - an extract of Cyc that represents the common sense and vetted relationships amongst KBpedia's base 55,000 concepts;
- DBpedia - a machine-readable version of parts of Wikipedia in RDF;
- GeoNames - a geographical database of some 10 million places linked to about 800 distinct feature classes; and
- UMBEL - the initial organizational structure for the knowledge graph.

We have mapped and re-expressed each of these sources into the single, coherent knowledge system of KBpedia. We split the resulting KBpedia knowledge graph along the lines of concepts and topics, entities, events, attributes, annotations, and relations and their associated natural classifications or types. This resulting combination gives KBpedia a rich set of structural components.

Wikipedia and Wikidata are the two most important KB sources, Wikipedia for concepts, Wikidata for instances and properties, and both for multi-lingual capabilities. Certain aspects of Wikipedia have proven their usefulness for general knowledge acquisition, for example, using article (concept or entity) content to inform topical tagging using explicit semantic analysis (ESA),[4] automatic topic identification,[5] information extraction[6] or a myriad of others. A weakness of Wikipedia has been its category structure, which was not part of the original design but added in 2004. Various reviewers have likened Wikipedia more to a thesaurus than a classification scheme, others that it is different from classical knowledge organization systems in that it has no specified root or hierarchy. This situation improved a wee bit from 2006 to 2010, when editors organized the main Wikipedia topics according to top-level and main topics.[7] Still, typical commentaries point to the fact that Wikipedia's category structure is "noisy, ill-formed, and difficult to make sense of."[8] Its crowd-sourced nature has led to various direct and indirect cycles in portions of the cate-

gory structure. All of these problems lead to the inability to do traditional reasoning or inference over the Wikipedia category graph.[9] We have done much to clean the Wikipedia categories and remapping them and their instances to KBpedia has now made the structure computable.

The choice of Wikipedia's founders to make its full content available electronically for free and without restriction was a masterstroke, and now carries over to Wikidata, a sister project to Wikipedia under the Wikimedia banner. I only hope we can honor this philosophy. Wikidata takes as its starting point the structured data about entities evident in Wikipedia infoboxes. Rather than extracting and cleaning that entity information as DBpedia does, the roles of Wikidata are as a standalone reference base and as a multilingual source for all entities feeding the Wikimedia network, including Wikipedia. As of June 2018, Wikidata contained about 50 million data items in its system. The Wikidata approach leads to more uniformity and consistency and provides a central Wikimedia access point for structured data.[10] However, somewhat akin to Wikipedia, Wikidata also has struggled to find an appropriate typology (or ontology) for its millions of entities.[11] Again, KBpedia provides one such structure.

Besides these main six KBs, KBpedia has extended mappings to a further 20 other vocabularies, including schema.org, Dublin Core, the Bibliographic Ontology (BIBO), and others. KBpedia also supports exports in various formats and finite state transducers or specialty lists, used as inputs to third-party analysis, management, and visualization tools. We also transform external and domain data into KBpedia's internal canonical forms for interacting with the overall structure.

The KBpedia Typologies

A *typology* is a grouping of similar types, sharing some essential characters. Each type is a parent to a particular group of instances, which also share essential traits or attributes. *Chapter 10* is devoted to a discussion of KBpedia's typologies and the advantages of their modular, expandable design. *Table 10-2* provides a listing of the current typologies; *Table 10-3* describe those that are core.

STRUCTURE

This section goes into a bit more detail on the structure of the KBpedia knowledge graph, KKO. At each level in the KBpedia Knowledge Ontology, we strive to organize each entry according to Peirce's universal categories of Firstness (1ns), Secondness (2ns) and Thirdness (3ns). Most of the reference concepts (RCs) in KKO are organized under the Generals (3ns) branch, though that is not evident from inspection of the upper nodes alone. All of KKO's SuperTypes (typologies) reside there.

Most of the 30 or so core typologies in KBpedia do not overlap with one another, what is known as disjoint. Disjointness enables us to perform powerful reasoning and subset selection (filtering) on the KKO graph. Upper typologies are useful to organize core entities, plus they providing homes for shared concepts. Living Things, for ex-

ample, can capture concepts shared by all plants and animals, by all life, which then enables better segregation of those life forms. We apply such natural segregations across the KKO structure. Here is the upper structure of KKO with its 171 concepts:

Monads [1ns]
- **FirstMonads** [1ns]
  - **Suchness** [1ns]
    - **Accidental** [1ns]
    - **Inherent** [2ns]
    - **Relational** [3ns]
  - **Thisness** [2ns]
    - **Chance** [1ns]
    - **Being** [2ns]
    - **Form** [3ns]
  - **Pluralness** [3ns]
    - **Absolute** [1ns]
      - **Inclusive** [1ns]
      - **Exclusive** [2ns]
      - **Difference** [3ns]
    - **SimpleRelative** [2ns]
    - **Conjugative** [3ns]
- **DyadicMonads** [2ns]
  - **Attributives** [1ns]
    - **Oneness** [1ns]
      - **Identity** [1ns]
      - **Real** [2ns]
        - **Matter** [1ns]
        - **SubstantialForm** [2ns]
        - **AccidentalForm** [3ns]
      - **Fictional** [3ns]
    - **Otherness** [2ns]
    - **Inherence** [3ns]
      - **Quality** [1ns]
      - **Negation** [2ns]
      - **Intrinsic** [3ns]
  - **Relatives** [2ns]
    - **Concurrents** [1ns]
    - **Opponents** [2ns]
    - **Conjunctives** [3ns]

Quantity [1ns]
Values [1ns]
Numbers [1ns]
Multitudes [2ns]
Magnitudes [3ns]
Discrete [2ns]
Complex [3ns]
Subsumption [2ns]
Connective [3ns]
Unary [1ns]
Binary [2ns]
Conditional [3ns]
Indicatives [3ns]
Iconic [1ns]
Indexical [2ns]
Associative [3ns]
Denotative [1ns]
Similarity [2ns]
Contiguity [3ns]
TriadicMonads [3ns]
Representation [1ns]
Icon [1ns]
Index [2ns]
Symbol [3ns]
Mediation [2ns]
Mentation [3ns]
Particulars [2ns]
MonadicDyads [1ns]
MonoidalDyad [1ns]
EssentialDyad [2ns]
InherentialDyad [3ns]
Events [2ns]
Spontaneous [1ns]
Action [2ns]
Exertion [1ns]
Perception [2ns]
Thought [3ns]
Continuous [3ns]
TriadicAction [1ns]
Activities [2ns]

Processes [3ns]
Entities [3ns]
SingleEntities [1ns]
Phenomenal [1ns]
States [2ns]
Situations
Continuants [3ns]
Space
Points [1ns]
Areas [2ns]
2D-dimensions
SpaceRegions [3ns]
3D-dimensions
Time
Instants [1ns]
Intervals [2ns]
Eternals [3ns]
PartOfEntities [2ns]
Members [1ns]
Parts [2ns]
FunctionalComponents [3ns]
ComplexEntities [3ns]
CollectiveStuff [1ns]
MixedStuff [2ns]
CompoundEntities [3ns]
Generals [3ns] (== SuperTypes)
Constituents [1ns]
NaturalPhenomena [1ns]
SpaceTypes [2ns]
Shapes [1ns]
Places [2ns]
LocationPlace
AreaRegion
Forms [3ns]
TimeTypes [3ns]
Times [1ns]
EventTypes [2ns]
ActivityTypes [3ns]
Predications [2ns]
AttributeTypes [1ns]

- **IntrinsicAttributes** [1ns]
- **AdjunctualAttributes** [2ns]
- **ContextualAttributes** [3ns]

RelationTypes [2ns]

- **DirectRelations** [1ns]
- **CopulativeRelations** [2ns]
  - ActionTypes
- **MediativeRelations** [3ns]
  - SituationTypes

RepresentationTypes [3ns]

- **Denotatives** [1ns]
- **Indexes** [2ns]
- **Associatives** [3ns]

Manifestations [3ns]

NaturalMatter [1ns]

- **AtomsElements** [1ns]
- **NaturalSubstances** [2ns]
- **Chemistry** [3ns]

OrganicMatter [2ns]

- **OrganicChemistry** [1ns]
  - BiologicalProcesses
- **LivingThings** [2ns]
  - **Prokaryotes** [1ns]
  - **Eukaryotes** [2ns]
    - **ProtistsFungus** [1ns]
    - **Plants** [2ns]
    - **Animals** [3ns]
  - **Diseases** [3ns]
- **Agents** [3ns]
  - **Persons** [1ns]
  - **Organizations** [2ns]
  - **Geopolitical** [3ns]

Symbolic [3ns]

- **Information** [1ns]
  - **AVInfo** [1ns]
    - VisualInfo
    - AudioInfo
  - **WrittenInfo** [2ns]
  - **StructuredInfo** [3ns]
- **Artifacts** [2ns]

	FoodDrink
	Drugs
	Products
	Facilities
Systems [3ns]	
	ConceptualSystems [1ns]
	Concepts [1ns]
	TopicsCategories [2ns]
	LearningProcesses [3ns]
	SocialSystems [2ns]
	FinanceEconomy
	Society
	Methodeutic [3ns]
	InquiryMethods [1ns]
	KnowledgeDomains [2ns]
	EmergentKnowledge [3ns]

Table B-1: The KKO Upper Structure Organized by the Universal Categories

Note that *Table 10-2* in the main book provides an expansion on the typologies found under the Generals branch in the table above.

CAPABILITIES AND USES

Online demos, various search and discovery facilities, and documentation on the KBpedia Web site provide further details about KBpedia. The primary purpose of KBpedia is to serve as a starting template for creating local domain knowledge graphs. However, as is, KBpedia also has the following capabilities:

- A consistent, coherent combination of six (6) large and leading public knowledge bases into the computable KBpedia knowledge structure;
- Mappings to a further 20 'extended' knowledge bases;
- A structured organization of the contributing knowledge sources that enables separate treatment of concepts, entities, events, attributes, and relations and their associated types;
- Powerful and flexible manipulation and filtering capabilities;
- Robust and configurable search and retrieval functions;
- Pre-built taggers, classifiers, and mappers;
- Ingest and export of multiple data formats;
- All functions available via microservice-like APIs and Web services;

- Use of open and accepted languages and standards;
- A modular and expandable architecture;
- A completely Web-based system, which we may deploy locally or in the cloud;
- Integration and incorporation of all data assets — unstructured text, semi-structured and markup data, and structured datasets and databases;
- A reference structure for inter-relating and integrating your domain content;
- Inherent multi-linguality, supported by the 200+ languages of the source knowledge bases;
- Precise semantic representation for all items, enabling better selections and matches;
- The ability to make selections via inference and other logical operations;
- The potential to recognize and train up to 47,000 fine-grained entity types;
- A knowledge graph suitable for network analytics such as influence, centrality, shortest paths, assortative mixing, and betweenness;
- Faster, cheaper creation of positive and negative machine learning training sets; and
- Faster, cheaper configuration and testing of machine learners.

You may download the open source KBpedia and other supporting materials and documentation from the project's GitHub site.

Appendix Notes

1. Material in this appendix is drawn with permission from the KBpedia Web site at http://kbpedia.org.
2. This number is as of version 1.60 of KBpedia, based on the completion of this book in the first half of 2018. The current publicly released version of KBpedia likely has a different number of reference concepts. Back-of-the-envelop estimates suggest KBpedia should eventually grow to be on the order of 80,000 reference types.
3. Bergman, M. K., "Cognonto is on the Hunt for Big AI Game," *AI3:::Adaptive Information*, Sep. 2016.
4. Gabrilovich, E., and Markovitch, S., "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis," *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India: 2007.
5. Hassan, M., "Automatic Document Topic Identification Using Hierarchical Ontology Extracted from Human Background Knowledge," 2013.
6. Wu, F., and Weld, D. S., "Automatically Refining the Wikipedia Infobox Ontology," *Proceedings of the 17th International Conference on World Wide Web*, ACM, 2008, pp. 635–644.
7. For Wikipedia's main topics, see http://en.wikipedia.org/wiki/Category:Main_topic_classificationsReference; for Wikipedia's top-level categories, see http://en.wikipedia.org/wiki/Category:Fundamental_categories.
8. Kittur, A., Chi, E. H. an. S., and B., "What's in Wikipedia? Mapping Topics and Conflict Using Socially Annotated Category Structure," *Proceedings of the 27th Annual CHI Conference on Human Factors in Computing Systems*

(CHI'2009), New York, USA: 2009, pp. 1509–1512.

9. Paulheim, H., and Bizer, C., "Type Inference on Noisy RDF Data," *International Semantic Web Conference*, Springer, 2013, pp. 510–525.
10. Vrandečić, D., and Krötzsch, M., "Wikidata: A Free Collaborative Knowledgebase," *Communications of the ACM*, vol. 57, 2014, pp. 78–85.
11. As of May 2018, Wikidata contained more than 54 million entities. However, there has yet to emerge an overarching typology or ontology for these entities, with the typing system that does exist growing from the bottom up. For some background, see https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Migrating_away_from_GND_main_type.