

A Knowledge Representation Practionary

Michael K. Bergman

A Knowledge Representation Practitioner's

Guidelines Based on Charles Sanders Peirce

 Springer

Michael K. Bergman
Cognonto Corporation
Coralville, IA, USA

ISBN 978-3-319-98091-1 ISBN 978-3-319-98092-8 (eBook)
<https://doi.org/10.1007/978-3-319-98092-8>

Library of Congress Control Number: 2018954477

© Springer Nature Switzerland AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*To
Wendy*

Preface

Human language is not the starting point for knowledge representation. Our utterances or our symbols are not the basis for what we desire to convey; they are only representations. Knowledge, the actionable side of information, is rooted in something more fundamental than language. What that something may be is what this book is about.

Competing factions have claimed truth since at least the beginning of communication. Who knows, maybe bees, whales, dingoes, and apes also have communities believing different things as true, perhaps even leading to conflict. As humans, we know from wars, missed opportunities, and personal misunderstandings the tragedy that different premises of truth may bring. We have to admit if we want to represent human knowledge to computers that we humans have not done such a hot job representing knowledge to ourselves. Since we are starting out on a journey here to explore knowledge representation (KR) for knowledge management, artificial intelligence, and other purposes, more than a bit of humility seems in order.

Information, by no means a uniformly understood concept, arises from a broader context than gestures, symbols, or sounds. For some, information is energy or when missing is entropy, the nuts-and-bits of messages. For some, information is meaning. That we continue to use ‘information’ in these senses, and more, in fact, tells us these senses are properly within the boundaries of the concept. Still, even if we can clear the hurdle of grokking information, we have the next obstacle of deciphering what is knowledge, that which next lies directly on our path. Further, of course, we then need to record somehow and convey all of this if we are to represent the knowledge we have gained to others. Like I say, if we have a hard time communicating all of this to other humans, what can we say about our ability to do so to machines and AI?

But maybe I overthink this. Any tasks us humans do using information that we can automate with acceptable performance may lead to more efficiency and perhaps more job satisfaction for the workers involved, maybe even more wealth. Conversely, maybe this automation leads to loss of jobs for the workers. I do know, however, if we are ever to rely upon machines to work on our behalf, requiring little or no oversight, then we need to figure out what this knowledge is and how to represent it to the machine. Such is the task of KR. What I try to provide in this book is a way to

think and a practical guidebook of sorts for how to approach the questions of computers and knowledge.

The world is real. It exists independent of us or how we may think about it, though our thoughts are also part of our reality. Human history fills but a small thimble yet through the application of reason and truth-testing, including, since the Enlightenment, the scientific method, we humans have increasingly unveiled the truths of Nature, in the process creating wealth and comfort never before seen. Artificial intelligence (AI) will undoubtedly accelerate this trend. How fast that acceleration occurs is, in part, a function of how good we get at representing our knowledge. These representations are the encodings by which intelligent machines will work on our behalf. My quest in this treatise is to help promote this trend. I believe this quest to be noble and, in any case, inevitable. I believe there is something in our nature that compels us to pursue the path of useful information leading to knowledge.

The past decade was a golden one in advances in AI. We can now voice commands and requests to our phones and devices acting as virtual assistants. We are on the verge of self-driving vehicles and automation of routine knowledge worker tasks. Still, the deep learning that underlies many of these advances is an opaque, black box of indecipherable inferences. We don't know why some of this magic works or what the representations are upon which machines draw these inferences. For further advances to occur, for general AI or cognition to arise in silico, I believe we will need better ways to represent knowledge, reflective of the nature of information and its integral role in the real world.

I have had a passion for the nature and role of information throughout my professional life. I originally trained as an evolutionary biologist and population geneticist. Since my graduate days, I have replaced my focus on biological information with one based on digital information and computers. My passion has been on the role of information—biological or cultural—to confer adaptive advantage to deal with an uncertain future and as a means of generating economic wealth. My intuition—really, my underlying belief—is that there are commonalities between biological and cultural information. I have been seeking insights into this intuition for decades.

One of my first forays into information technology was a [data warehousing](#) venture, where the idea was to find ways to connect structured databases that, in native form, were stand-alone and unconnected. This venture coincided with the explosive growth of the initial Internet. To support the exploding content, we observed that large content suppliers were populating their web sites with searchable, dynamic databases, hidden from the search engines of that time (before Google's inception). We named this phenomenon the '[deep web](#)' and did much to define its huge extent and figure out ways to mine it. We saw that, in aggregate, the web was becoming a giant, global data warehouse, though largely populated by text content and less so by structured data. We shifted our venture emphasis to text and discovery. This shift raised the perplexing question of how to place information in text onto a common, equal basis to the information in a database, such as a structured record. (Yeah, I know, kind of a weird question.)

Tim Berners-Lee, inventor of the [World Wide Web](#), and colleagues put forward a vision of the [Semantic Web](#) in a *Scientific American* article in 2000.¹ The article painted a picture of globally interconnected data leveraged by agents or bots designed to make our lives easier and more automated. The late [Douglas Adams](#), of *Doctor Who* and *The Hitchhiker's Guide to the Galaxy* fame, had presciently produced a fascinating and entertaining TV program on the same topic for BBC2 about 10 years earlier. Called [Hyperland](#), you can see this self-labeled 'fantasy documentary' from 1990 in its entirety on [YouTube](#). The 50-min presentation, written by and starring Adams as the protagonist having a fantasy dream, features Tom, the semantic simulacrum (actually, [Tom Baker](#) from *Doctor Who*). Tom is the "obsequious, and fully customizable" personal software agent who introduces, anticipates, and guides Adams through what is a Semantic Web of interconnected information. Laptops (actually an early Apple), pointing devices, icons, and avatars sprinkle this *tour de force* in an uncanny glimpse into the (now) future.

One of the premises of the Semantic Web is to place what we now call unstructured, semi-structured, and structured information onto a common footing. The approach uses the [RDF](#) (Resource Description Framework) data model. RDF provided an answer to my question of how to combine data with text. I am sure there were other data models out there at the time that could have perhaps given me the way forward, but I did not discover them. It took RDF and its basic *subject-predicate-object* (*s-p-o*) 'triple' assertion to show me the way ahead. It was not only a light going on once I understood but the opening of a door to a whole new world of thinking about knowledge representation.

The usefulness of ideas behind the Semantic Web and the semantic technologies supporting it lured me to switch emphasis again. I founded a new company with [Frédéric Giasson](#), and we proceeded to provide semantic technology solutions to enterprises over the next 10 years. The Web today is almost unrecognizable from the Web of 15 years ago. If one assumes that Web technologies tend to have a 5-year or so period of turnover, we have gone through three to four generations of change on the Web since the initial vision for the Semantic Web.

Many of our engagements were proprietary, though we did provide three notable open source projects. We developed a general semantic platform for ontology (knowledge graph) and data management, the still-active [Open Semantic Framework](#) project. To help information interoperate, we created [UMBEL](#), a subset of Cyc and a contributor to our current efforts, as a set of reference concepts that users can share across different Web datasets. Based on that experience, we designed a successor reference knowledge structure, [KBpedia](#), a combination of upper knowledge graph and leading public knowledge bases. We talk much about KBpedia throughout since it is this book's reference knowledge structure.

The marrying of electronic Web knowledge bases—such as [Wikipedia](#) or internal ones like the Google search index or its [Knowledge Graph](#)—with improvements in [machine learning algorithms](#) is systematically mowing down what used to be called

¹ Berners-Lee, T., Lassila, O., and Hendler, J., "The Semantic Web," *Scientific American Magazine*, 2001.

the [Grand Challenges](#) of computing, such as machine translation or language understanding. Sensors are also now entering the picture, from our phones to our homes and our cars, that exposes the higher-order requirement for data integration combined with semantics. Natural language processing (NLP) kits have improved in accuracy and execution speed; many semantic tasks such as tagging or categorizing or questioning already perform at acceptable levels for most projects. We naturally call the marriage of these knowledge sources with AI ‘knowledge-based artificial intelligence.’ KBAI is one of the potential payoffs that would arise from better ways to represent knowledge and thus is a common theme throughout the book.

Combining information goes beyond the technical challenges of matching forms and formats. We need to tackle the question of meaning, inextricably entwined with context and perspective. Cinemaphiles will readily recognize [Akira Kurosawa’s *Rashomon*](#) film of 1951. In the 1960s, one of the most popular book series was [Lawrence Durrell’s *The Alexandria Quartet*](#). Both, each in its way, tried to get at the question of what is the truth by telling the same story from the perspective of different protagonists. Whether you saw Kurosawa’s movie or read Durrell’s books, you know the punchline: truth is very different depending on the point of view and experience—including self-interest and delusion—of each protagonist.

All of us recognize this phenomenon of the [blind man’s view of the elephant](#). The problem we are trying to solve is how to connect information meaningfully. For that, we need to somehow capture the ideas of perspective and context, as well as the usual vagaries of imprecise semantics. [Root cause analysis](#) for what it takes to achieve meaningful, interoperable information suggests one pivotal factor is to describe source content adequately in context to its use. Capturing and reflecting context is essential if we are to get information sources to work together, a capability we give the fancy label of ‘interoperability.’ We also need to assemble and represent this information such that we can reason over it and test new knowledge against it, a structural form we call a ‘knowledge graph.’ All of this requires a logical and coherent theory—a *grounding*—for how to represent knowledge.

Our client efforts over the past decade were converging on design thoughts about the nature of information and how to signify and communicate it. The bases of an overall philosophy regarding our work emerged around the teachings of [Charles Sanders Peirce](#) and [Claude Shannon](#), each explicating one of the boundary senses of information. Shannon emphasized the message and mechanical aspects of information; Peirce emphasized meaning in both breadth and depth. In the combination, we see semantics and groundings as essential to convey accurate messages. Simple forms, so long as they are correct, are always preferred over complex ones because message transmittal is more efficient and less subject to losses (inaccuracies). How we could represent these structures in graphs affirmed the structural correctness of our design approach. The now visible reawakening of artificial intelligence helps to put the Semantic Web in its proper place: a key subpart, but still a subset, of AI.

I first encountered Charles S. Peirce from the writings of [John Sowa](#) about a decade ago. Sowa’s writings are an excellent starting point for learning about logic

and ontologies, especially his articles on Peirce and signs.² Early on it was clear to me that knowledge modeling needed to focus on the inherent meaning of things and concepts, not their surface forms and labels. Sowa helped pique my interest that Peirce's [theory of semiotics](#) was perhaps the right basis for getting at these ideas.

In the decade since that first encounter, I have based some writings on Peirce's insights. I have developed a fascination with his life and teachings and thoughts on many topics. I have become convinced that Peirce—an American philosopher, logician, scientist, and mathematician—was possibly one of the greatest thinkers ever. While the current renaissance in artificial intelligence can certainly point to the seminal contributions of [George Boole](#), Shannon, [Alan Turing](#), and [John von Neumann](#) in computing and information theory (among many others), my view, not alone, is that C.S. Peirce belongs in those ranks from the perspective of [knowledge representation](#), the *meaning* of information, and hewing to reality.

The importance of studying Peirce for me has been to tease out those principles, design bases, and mindsets that can apply Peircean thinking to the modern challenge of knowledge representation. This knowledge representation is like Peirce's categorization of science or signs but is broader still in needing to capture the nature of relations and attributes and how they become building blocks to predicates and assertions. In turn, we need to subject these constructs to logical tests to provide a defensible basis for what is knowledge and truth given the current information. Then, all of these representations need to be put forward in a manner (symbolic representation) that is machine readable and computable.

In reading and studying Peirce for more than a decade, it has become clear that he had insights and guidance on every single aspect of this broader KR problem. My objective has been to take these piece parts (Peirce parts?) and recombine them into a whole consistent with Peirce's *architectonic*. How can Peirce's thinking be decomposed into its most primitive assumptions to build up a new KR representation? These are the points I argue in the book while also sharing the experience of how we may integrate these viewpoints into working knowledge management systems.

I have no intent for balance in this exposition. There are wonderful textbooks and handbooks available if you are seeking a neutral presentation on knowledge representation in computer and information science. The lens I use is strictly that of Peirce and his views that contribute to an understanding of knowledge representation, at least how I read and understand those views. Peirce further guides the scope and organization of this book. One of Peirce's signal contributions was the philosophy of pragmatism, according to a specific maxim and a recommended methodology to follow, what the Peirce scholar Kelly Parker calls a 'practionary.' To my knowledge, this book employs this Peircean methodology for the first time. Given this emphasis, we will by necessity need to tackle many Peircean concepts, some with arcane or jaw-breaking labels. That is a small price to pay to gain entry into Peirce's brilliant insights.

I also minimize math and equations in the book. I provide many salient references for exploring topics further. I try to emphasize how to think and organize.

² Use [https://www.google.com/search?as_q="peirce"&as_sitesearch=jfsowa.com](https://www.google.com/search?as_q=) for a listing.

I avoid cookbook steps or prescriptive techniques or methods. I do not recommend specific tools. Rather, because of the coherence of Peirce's views, I use how I understand him and his writings, including interpretations by others, to bring a consistent approach, logic, and mindset to the question of knowledge representation. By straddling today's two separate worlds of Peirce scholarship and knowledge representation, I perhaps risk disappointing both camps. One of my points, though, is that the camps should be separated no longer.

I would first like to thank my colleague and business partner, Frédéric Giasson, for his creativity and effort in our commercial ventures over the past decade. He was not only the implementer of the many systems we developed, and a constant fount of ideas and innovation, but a great friend and a calm and cool influence during those engagements. Though I am the recorder of the results in this book, he deserves co-billing for why and how this book came into being.

I want to thank those who have encouraged me over many years to write this book, including from many commenters on my *A13::Adaptive Information* blog. I especially thank Fred, Steve Ardire, Alianna Maren, Alan Morrison, Amit Sheth, and Peter Yim for their encouragement. I further thank Amit for his kind efforts to help me find and secure a publisher.

I thank my former colleague, Jacque Bokow, for early editorial assistance and advice. I much appreciate the complete and detailed reviews I got on the first draft from Michael Buckland, Scott David, Rob Hillard, John Huntley, and Jack Park. I am grateful for the commentary and errors found in my readings of Peirce from Jon Alan Schmidt and Edwina Taborsky, as well as insights I have gained from the Peirce-L discussion group. I further thank William Anderson, Andreas Blumauer, Fred Giasson, Alan Morrison, Amit Sheth, Aleksander Smywiński-Pohl, Bobbin Teagarden, and Tom Tiaht for their reviews and commentary. Despite their best efforts to find and correct my errors and to make great suggestions, I am sure that errors remain, which are entirely my responsibility. I ask your forbearance for any errors or oversights. I lastly thank Susan Lagerstrom-Fife and Caroline Flanagan for helping to shepherd the manuscript through the publication process.

I find it wondrous that the human species has come to learn and master symbols. That mastery, in turn, has broken the shackles of organic evolution and has put into our hands and minds the very means and structure of information itself. The *lingua franca* for doing so is knowledge representation, best done, I believe, following the guidelines of Charles Sanders Peirce.

Coralville, IA
October 2018

Michael K. Bergman

In-line Citations

Here are the conventions and sources used for quotations for Peirce’s writings as used in the book. Items separated by a period or colon are page or clause number.

Abbrev.	Example	Source
CP	CP 1.343	Peirce, C. S., <i>The Collected Papers of Charles Sanders Peirce, 8 Volumes</i> , Cambridge, MA: Harvard University Press, 1931.
EP	EP 2:43	Peirce, C. S., <i>The Essential Peirce: Selected Philosophical Writings, Vol 1 (1867–1893)</i> , Bloomington: Indiana University Press, 1992. Peirce, C. S., <i>The Essential Peirce: Selected Philosophical Writings, Vol 2 (1893–1913)</i> , Bloomington: Indiana University Press, 1998.
MS	MS 32	Robin, R. S., <i>Annotated Catalogue of the Papers of Charles S. Peirce</i> , Amherst, Massachusetts: The University of Massachusetts Press, 1967. Robin, R. S., “The Peirce Papers: A Supplementary Catalogue,” <i>Transactions of the Charles S. Peirce Society</i> , 1971, pp. 37–57.
NEM	NEM 4:83	Peirce, C. S., <i>The New Elements of Mathematics by Charles S. Peirce</i> , Hague, Netherlands: The, Mouton Publishers, 1976 (four volumes).
W	W 3:266	Peirce, C. S., <i>The Writings of Charles S. Peirce—Chronological Editions Vol 8</i> , compiler, Peirce Edition Project, Bloomington: Indiana University Press, 1982–2009 (by volume).

I have attempted to date each Peirce quote, given the current tendency in scholarship and its usefulness to place his views into a chronology. Unlike most practice, I list the year first and then the citation.

Permission

Significant portions of the material in this book were first published on the author’s *AI3::Adaptive Information* blog, at <http://mkbergman.com>. We thank the author for the permission to use this copyrighted material.

Contents

1	Introduction	1
	Structure of the Book	2
	Overview of Contents	3
	Key Themes	10
2	Information, Knowledge, Representation	15
	What Is Information?	16
	What Is Knowledge?	27
	What Is Representation?	33
Part I Knowledge Representation in Context		
3	The Situation	45
	Information and Economic Wealth	46
	Untapped Information Assets	54
	Impediments to Information Sharing	61
4	The Opportunity	65
	KM and a Spectrum of Applications	66
	Data Interoperability	69
	Knowledge-Based Artificial Intelligence	74
5	The Precepts	85
	Equal Class Data Citizens	86
	Addressing Semantic Heterogeneity	91
	Carving Nature at the Joints	97
Part II A Grammar for Knowledge Representation		
6	The Universal Categories	107
	A Foundational Mindset	108
	Firstness, Secondness, Thirdness	112
	The Lens of the Universal Categories	116

7 A KR Terminology 129
 Things of the World 131
 Hierarchies in Knowledge Representation 135
 A Three-Relation Model 143

8 KR Vocabulary and Languages 151
 Logical Considerations 153
 Pragmatic Model and Language Choices 163
 The KBpedia Vocabulary 167

Part III Components of Knowledge Representation

9 Keeping the Design Open 183
 The Context of Openness 184
 Information Management Concepts 193
 Taming a Bestiary of Data Structs 200

10 Modular, Expandable Typologies 207
 Types as Organizing Constructs 208
 A Flexible Typology Design 215
 KBpedia’s Typologies 219

11 Knowledge Graphs and Bases 227
 Graphs and Connectivity 228
 Upper, Domain, and Administrative Ontologies 237
 KBpedia’s Knowledge Bases 242

Part IV Building KR Systems

12 Platforms and Knowledge Management. 251
 Uses and Work Splits 252
 Platform Considerations 262
 A Web-Oriented Architecture 268

13 Building Out the System. 273
 Tailoring for Domain Uses 274
 Mapping Schema and Knowledge Bases 280
 Pay as You Benefit. 291

14 Testing and Best Practices 295
 A Primer on Knowledge Statistics 296
 Builds and Testing. 304
 Some Best Practices 309

Part V Practical Potentials and Outcomes

15 Potential Uses in Breadth 319
 Near-Term Potentials 320
 Logic and Representation 327
 Potential Methods and Applications 332

- 16 Potential Uses in Depth** 343
 - Workflows and BPM 343
 - Semantic Parsing 349
 - Cognitive Robotics and Agents 361
- 17 Conclusion** 371
 - The Sign and Information Theoretics 372
 - Peirce: The Philosopher of KR 373
 - Reasons to Question Premises 377
- Appendix A: Perspectives on Peirce** 381
- Appendix B: The KBpedia Resource** 409
- Appendix C: KBpedia Feature Possibilities** 421
- Glossary** 435
- Index** 451