



# Deep Web Annotation with Special-Purpose Ontologies



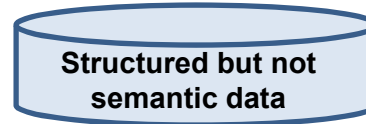
# How Much Data is Available?

External/public



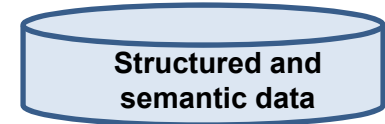
Unstructured data

25 billion documents  
at Google



Structured but not  
semantic data

782 APIs at  
ProgrammableWeb



Structured and  
semantic data

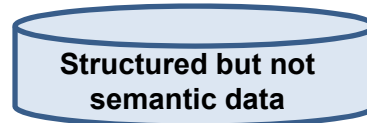
645 million triples\* at  
Swoogle

Internal/private



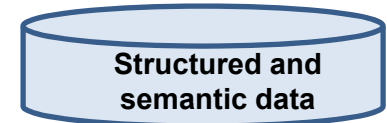
Unstructured data

E.g. documents on  
desktop



Structured but not  
semantic data

E.g. enterprise  
databases



Structured and  
semantic data

???



## ...And There is Much More: The Deep Web

“The Deep Web (also called Deepnet, the invisible Web, or the hidden Web) refers to World Wide Web content that is not part of the surface Web, which is indexed by search engines. [...] In 2000, it was estimated that the deep Web contained approximately 7,500 terabytes of data and 550 billion individual documents. Estimates – based on extrapolations from a study done at University of California, Berkeley – show that the deep Web consists of about **91,000 terabytes**. By contrast, the surface Web (which is easily reached by search engines) is only about 167 terabytes. The Library of Congress contains about 11 terabytes.”



# Current Situation\*: General-Purpose Ontologies and Deep Web Not Connected

General-purpose ontologies



Low value for business (no critical mass of instances)

The Deep Web

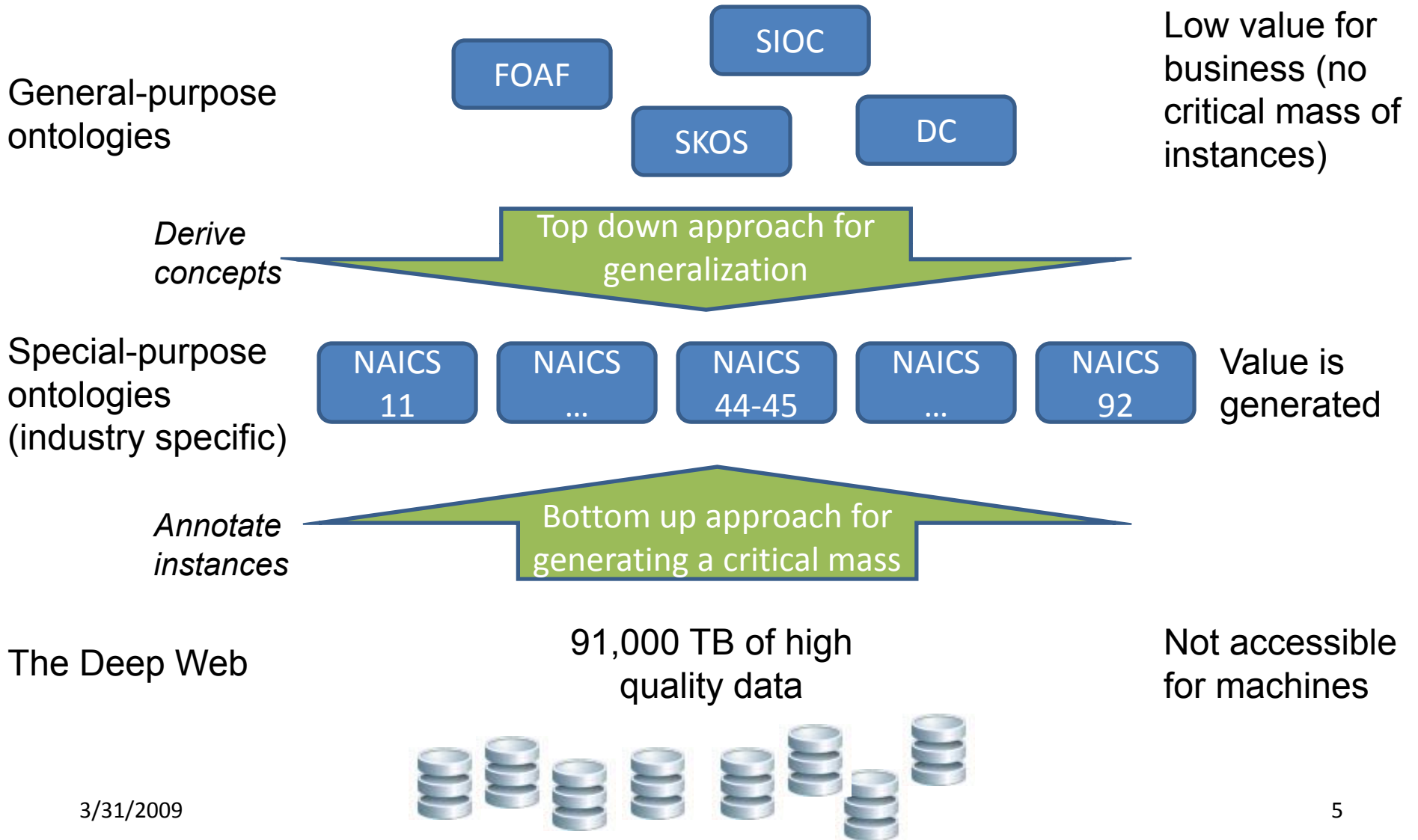
91,000 TB of high quality data

Not accessible for machines





# Annotate the Deep Web in Order to Generate a Critical Mass of Knowledge For Certain Service Systems





# Use Case: Retail Industry

