Big Data that might benefit from ontology technology, but why this usually fails

Barry Smith

National Center for Ontological Research

The strategy of annotation

Databases describe data using multiple heterogeneous labels

If we can *annotate* (tag) these labels using terms from common controlled vocabularies, then a virtual arms-length integration can be achieved, providing

- immediate benefits for search and retrieval
- a starting point for the creation of net-centric reference data
- potential longer term benefits for reasoning

with no need to modify existing systems, code or data

See Ceusters *et al. Proceedings of DILS 2004.*
http://ontology.buffalo.edu/bio/LinkSuite.pdf

String searches yield partial results, rest on manual effort and on familiarity with existing database contents

Ontologies facilitate grouping of annotations

brain	**20**
hindbrain	**15**
rhombomere	**10**

Query ‘brain’ without ontology 20
Query ‘brain’ with ontology 45

Examples of where this method works

- Reference Genome Annotation Project
  http://www.geneontology.org/GO.refgenome.shtml
- Human resources data in large organizations
  http://www.youtube.com/watch?v=OzW3Gc_yA9A
- Military intelligence data
  Salmen *et al.* in http://ceur-ws.org/Vol-808/

Other potential areas of application:

- Crime
- Insurance
- Public health
- Finance

But normally the method does not work

Semantic technology (OWL, …) seeks to break down data silos

Unfortunately it is now so easy to create ontologies that myriad incompatible ontologies are being created in *ad hoc* ways leading to the creation of new, semantic silos

The Semantic Web framework as currently conceived and governed by the W3C (modeled on html) yields minimal standardization

The more semantic technology is successful, they more we fail to achieve our goals

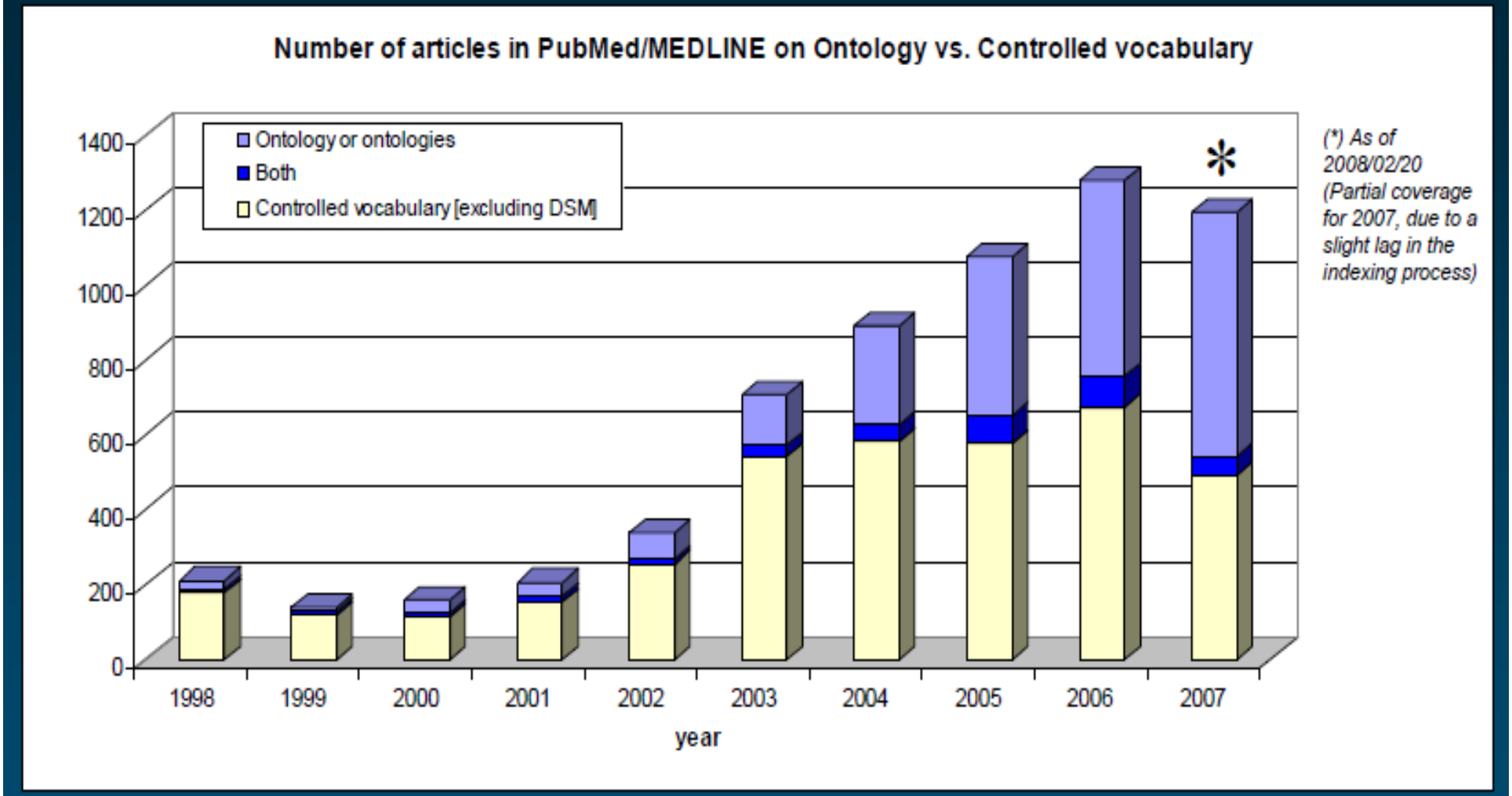
Reasons for this effect

- Just as it's easier to build a new database, so it's easier to build a new ontology for each new project
- You will not get paid for reusing existing ontologies (Let a million ontologies bloom)
- There are no 'good' ontologies, anyway (just arbitrary choices of terms and relations …)
- Information technology (hardware) changes constantly, not worth the effort of getting things right

How to do it right?

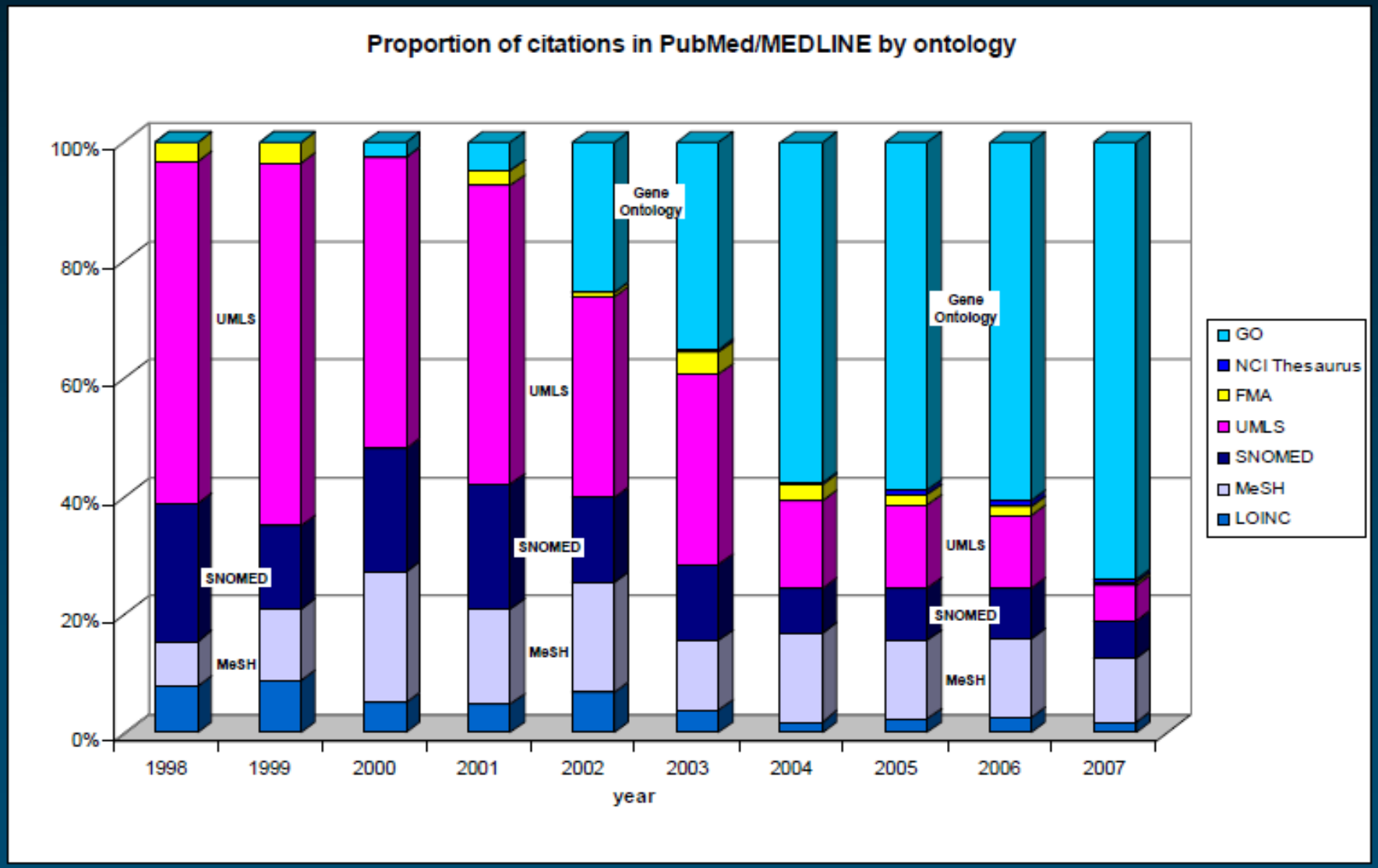
- how create an incremental, evolutionary process, where what is good survives, and what is bad fails
- create a scenario in which people will find it profitable to reuse ontologies, terminologies and coding systems which have been tried and tested
- silo effects will be avoided and results of investment in Semantic Technology will cumulate effectively

Biomedical ontology in PubMed



[Bodenreider, YBMI 2008]

By far the most successful: GO (Gene Ontology)



[Bodenreider, YBMI 2008]

GO provides a controlled vocabulary of terms for use in annotating (tagging) biological data

- multi-species, multi-disciplinary, open source
- built and maintained by domain experts
- contributing to the cumulativity of scientific results obtained by distinct research communities
- natural language and logical definitions for all terms to support consistent human application and computational exploitation
- rigorous governance process
- feedback loop connects users to editors

How to do it right

- ontologies should mimic the methodology used by the GO (following the principles of the OBO Foundry: http://obofoundry.org)
- ontologies in the same field should be developed in coordinated fashion to ensure that there is exactly one ontology for each subdomain
- ontologies should be developed incrementally in a way that builds on successful user testing at every stage