Ontology Evaluation: Methods and Metrics

MITRE Research Interest

Dr. Joanne S. Luciano

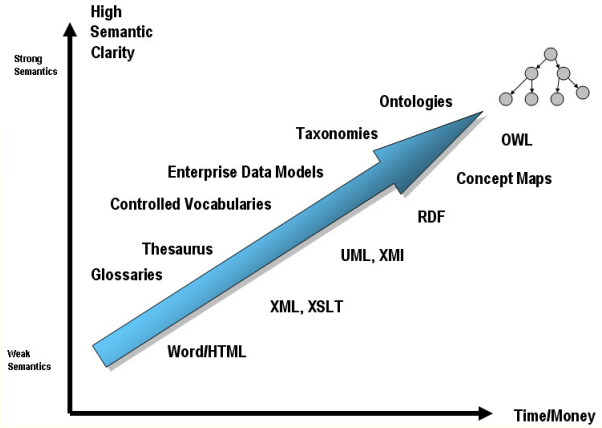
In collaboration with

Dr. Leo Obrst, PhD
Suzette Stoutenburg
Kevin Cohen
Jean Stanford

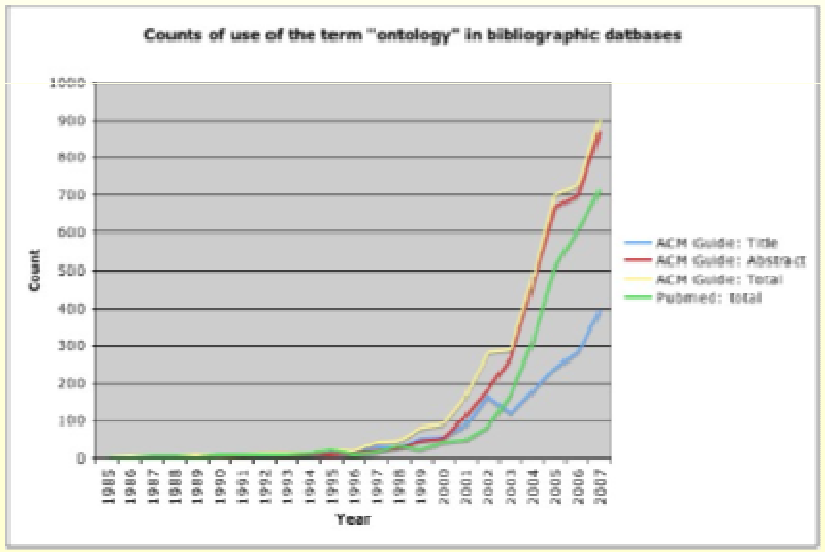
MITRE

Ontology: A Key Technology for Knowledge Management

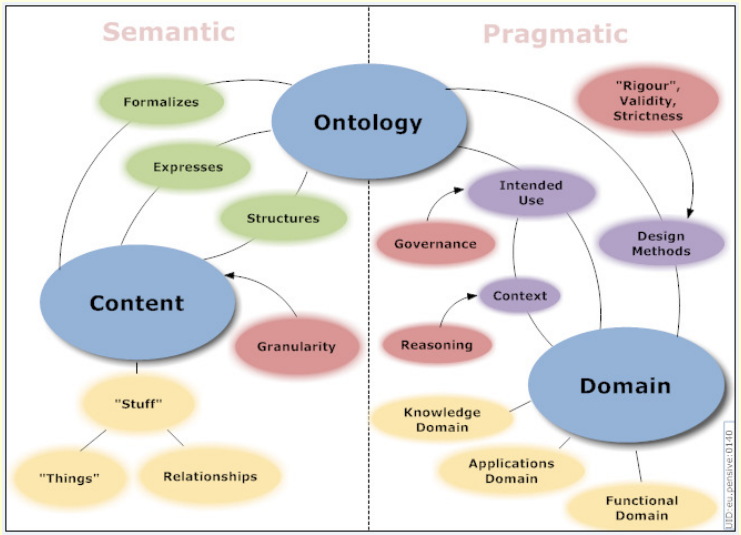
Used to describe terms in a vocabulary and the relationships among them. Ontology languages vary in their semantic expressiveness.



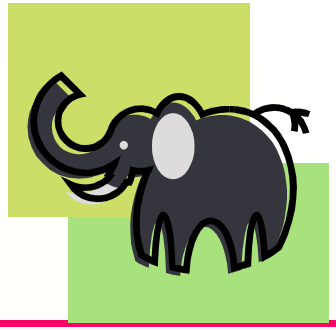
Based on work by Leo Obrst of MITRE as interpreted by Dan McCreary. This can be viewed as a trade-off of semantic clarity v. the time and money it takes to construct http://www.mkbergman.com/?m=20070516.



Ontologies have become the most widespread form of knowledge representation for multiple purposes

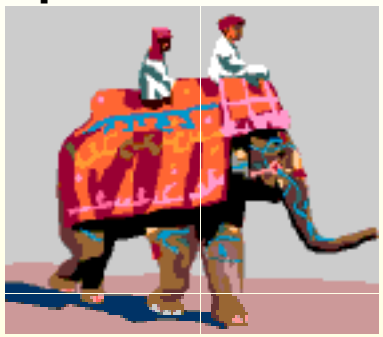


The Problem
Ontology Elephants

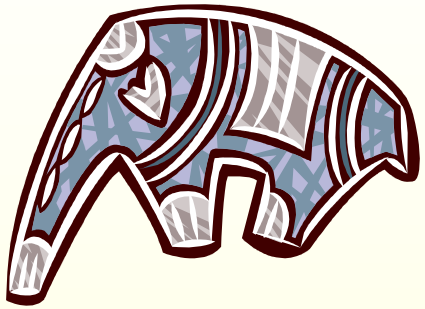


There is no single real elephant

There must be a purpose for an elephant: use cases?

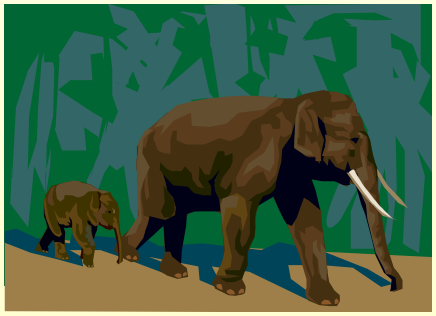


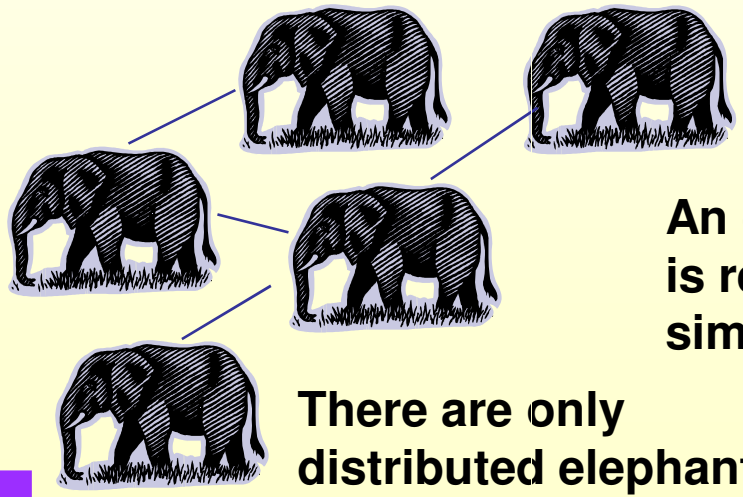
An elephant is abstract



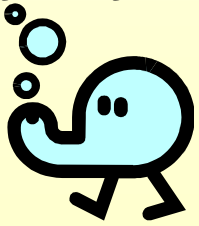
An elephant is very abstract

There must be an upper elephant





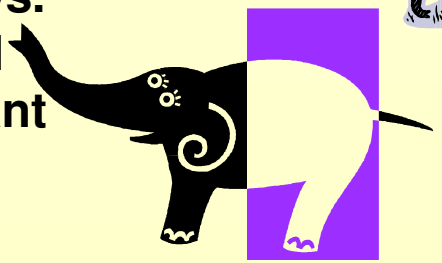
An elephant is really very simple



An elephant is the result of consensus

There are only distributed elephants & their mappings

Open vs. Closed Elephant



Prospects and Possibilities for Ontology Evaluation: The View from NCOR, Obrst, Hughes, Ray, WWW 2006, May 22–26, 2006, Edinburgh, UK.

MITRE

The Problem

Users need to be able to build sound ontologies and to reuse ontologies for different purposes. There is no standard no way to do that now.

At the Ontology Summit 2008 two competing "state of the art" evaluation proposals for the Open Ontology Repository were presented. Both treat ontologies as "black boxes" and both are subjective evaluations.

- Peer Review – self appointed editorial review board decides, non-overlapping domain so first one gets preference, 'best practices'
- User Ratings – users report their experience and rate them on a website

Language was inserted into the communiqué to provide mechanisms that enable ontology evaluation by other metrics

Prior work on evaluating ontologies is limited, no consensus

- Ontology Workshop Methods and Metrics October 2007

Workshop materials at: http://sites.google.com/a/cme.nist.gov/workshop-on-ontology-evaluation/

- Formal logic based applications can use software reasoners to address logical consistency and classification; many don't take advantage of these tools
- Natural Language Processing (NLP) applications use NLP evaluation methods, but address only NLP applications (text mark-up – aka "annotation"), information retrieval and extraction
- Alignment (mapping of ontologies) for data mining, integration, fusion

Ontology Summit 2007 (NIST, Gaithersburg, MD, April 23-24, 2007) See slide at end with notes.

Objective: Evolve toward Science & Engineering Discipline for Ontology

- Create procedures, processes, methods to help define, adjudicate, & ensure quality of knowledge capture/representation
- Facilitate the education of communities on ontology development & promote best practices for ontology development
- Enable the best standards related to ontologies, & promote linkages, liaisons among standards organizations

Approach:

- Two stages:
  - Recast use case into its components:
    - Functional objective
    - Design objective & requirements specification
    - Semantic components required to achieve above
  - Evaluate components using objective metrics

- Place existing evaluation methods in context by utility

- Engage and rally the community / stakeholders
  - Participate at appropriate meetings: present work and facilitate community focus on objective metrics for evaluation
  - Introduce users with complementary skills, joint vision, shared needs to develop needed metrics, content, tests, tools
  - Involve multiple government agencies, industry and academia to support initiative

Research Plan: (1) Identify use cases

For each, recast the use cases into their components:

a. Specify functional objective (what it is, what it does)

   e.g. enable investigation of data collected on influenza strain mutations that cause death in birds

b. Specify design objective (how good it has to be e.g., what specifications have to be met? Is it a prototype, for commercial use, or must it meet military specifications?)

   e.g. Must meet: *Minimum Information about an Influenza Genotype and a Nomenclature* Standard (MIIGNS)

c. Identify (or specify) the semantic components required to achieve the functional objective to the level specified by the design objective (Authoritative Sources such as engineering tolerances, physical constants, legal jurisdictions, company policies)

   e.g. To meet MIIGNS, the following semantic components must be included:

   biomaterial transformations

   assays

   data transformations

Research Plan (2) Develop Metrics

Develop metrics for 3 criteria for evaluation:

1. Correctness: how well the functional components express the design objectives
   a) Language expressiveness
   b) Fluency / competency
2. Completeness: combines use case with design criteria (requirement specification)
   a) To what extent can requirements be met?
   b) Which semantic components (authoritative sources) are needed/missing?
3. Utility: Is it useful? Does it work?

   Combine 1 and 2 (correctness and completeness) and evaluate against the use case (by competency questions, or other challenge tests).

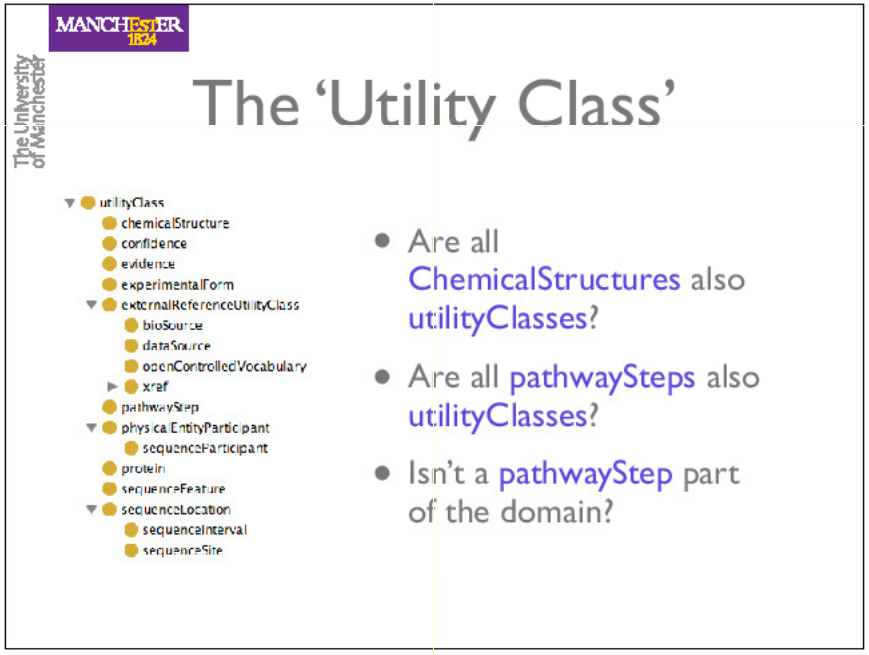
Examples:

- BioPAX (prior work)
- Habitat-Lite (subset of Environmental Ontology to support of NSF funded Mining Metadata for Metagenomics)
- Influenza Infectious Disease Ontology (for Genomics for Bioforensics MSR)

Example (1) BioPAX lack of fluency

chemical structure & pathway steps incorrectly modeled

- misunderstanding of the language (language has capability)
- modeled disjoint from the biology & chemistry
- leads to logical inconsistency



OWL has a steep learning curve, it's easy to get things wrong.

MITRE

Example (2) Habitat-Lite: correctness & completeness

Objective: facilitate capture of habitat and environmental metadata on genomic sequences

Approach: select subset of terms with highest frequency and evaluate usefulness by correctness and completeness metrics

- Evaluated correctness
  - 64% agreement (84 of 132 terms) of automated and expert mapping of terms
- Evaluated coverage of terms
  - 84% exact matches ("host," "aquatic," and "soil" covered 75%)

Hirschman, Clark, Cohen, Mardis, Luciano, Kottmann, Cole, Markowitz, Kyprpides, Field
Habitat-Lite: a GSC Case Study Based on Free Text Terms for Environmental Metadata
OMICS A Journal of Integrative Biology Volume 12, Number 2, 2008 (in press)

Example (3) Enable Influenza Research

(proposed construction and subsequent evaluation)

Function: enable investigation of data collected on influenza strain mutations that cause death in birds

Design objective: Minimum Information about an Influenza Genotype and a Nomenclature Standard (MIIGNS)

Semantic components:

1. **biomaterial transformations**
   - **a. recombinant plasmid biomaterial transformation**
   - **b. site-directed mutagenesis biomaterial transformation**
   - **c. reverse genetic virus production biomaterial transformation**
   - **d. Mouse infection biomaterial transformation**
2. **assays**
   - **a. weight assay**
   - **b. virus replication / mouse lung assay**
   - **c. Cytokine quantification assay**
3. **data transformations**
   - **a. statistical difference evaluation**

MITRE

Example (3) Enable Influenza Research

(proposed construction and subsequent evaluation)

Correctness:

- Language expressivity: validate definitions against OBO Foundry relations
- Fluency: inter-developer agreement (3 developers, 2 code same, 3rd validates)

Completeness:

- Calculate % coverage of minimum terms (18 terms)
- Calculate % coverage of full terms (196 terms)

Utility: Challenge Questions

- To be developed (by our collaborator BioHealthBase)

Impact

Communities of Practice areas need objective methods and Metrics to facilitate the development, interoperability and reuse of their ontologies

Some examples:

- Message Based Data Exchange
- BioSecurity
- Health Care and Biomedicine
- Life Sciences
- Disease
- Metagenomics
- Agile systems for rapid enterprise integration of heterogeneous data
- Intelligence Community

Why MITRE?

MITRE is uniquely positioned to act as an impartial experimental designer and arbitrator in the development of an evaluation methodology for ontologies

- MITRE has acted in the past for natural language technologies such as text summarization in the Text REtrieval Conferences (TREC) and Information Extraction in the BioCreAtIvE challenge

Additional Notes on Specific Slides

Slide 1: Lower right gaphic: Ontology Summit 2007 (NIST, Gaithersburg, MD, April 23-24, 2007) effort. See the following:
Ontology Summit 2007 - Ontology, Taxonomy, Folksonomy: Understanding the Distinctions. http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2007.
Ontology Summit 2007 Communique. http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2007_Communique.
Ontology Summit 2007 Ontology Dimensions Map. http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2007_FrameworksForConsideration/DimensionsMap.
Gruninger, Michael; Olivier Bodenreider; Frank Olken; Leo Obrst; Peter Yim. 2008. The 2007 Ontology Summit Joint Communiqué. Ontology, Taxonomy, Folksonomy: Understanding the Distinctions. Journal of Applied Ontology, forthcoming.

Slide 2: Ontology Summit 2008 (NIST, Gaithersburg, MD, April 28-29, 2008).
Ontology Summit 2008: Toward an Open Ontology Repository. http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2008.
Ontology Summit 2008 Communique. http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2008_Communique.

Slide 4:
Concerning:
"At the Ontology Summit 2008 two competing "state of the art" evaluation proposals for the Open Ontology Repository were presented. Both treat ontologies as "black boxes" and both are subjective evaluations.

–Peer Review –self appointed editorial review board decides, non-overlapping domain so first one gets preference, „best practices"
–User Ratings –users report their experience and rate them on a website

These points were made during the summit and discussed more fully during the Quality and Gatekeeping session of the Ontology Summit 2008 (NIST, Gaithersburg, MD, April 28-29, 2008.

Background References

The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge evaluation consists of a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain. BioCreative. (http://biocreative.sourceforge.net/)

The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. (http://trec.nist.gov/overview.html)

The Message Understanding Conferences (MUC) were initiated and financed by DARPA to encourage the development of new and better methods of information extraction. The character of this competition -- many concurrent research teams competing against one another -- necessitated the development of standards for evaluation, e.g. the adoption of recall and precision. (http://en.wikipedia.org/wiki/Message_Understanding_Conference)

Lenat, Douglas B. "CYC: a large-scale investment in knowledge infrastructure," *Communications of the ACM*, Volume 38 , Issue 11 (November 1995) Pages: 33 – 38.

Project Halo (http://www.projecthalo.com/), is a project funded by Paul Allen's Vulcan Ventures. The project was initially led by Oliver Roup and Noah Friedland but is currently led by Mark Greaves, a former DARPA Program Manager. Project Halo is an attempt to apply Artificial Intelligence techniques to the problem of producing a "digital Aristotle" that might serve as a mentor, providing comprehensive access to the world's knowledge (http://en.wikipedia.org/wiki/Project_Halo).

Ontoprise is a commerial software provider of ontology-based solutions. (http://www.ontoprise.de/content/index_eng.html)

Maedche, Alexander and Staab, Steffen. Ontology learning for the semantic web. Special Issue on Semantic Web. IEEE Intelligent Systems, 16(2):72-79, MAR 2001.

López, M. F.; Gómez-Pérez, A.; Sierra, J. P. & Sierra, A. P. Building a chemical ontology using Methontology and the OntologyDesign Environment IEEE Intelligent Systems and Their Applications, 1999, 14, 37-46

Oltramari, A.; Gangemi, A.; Guarino, N. & Masolo, C. Restructuring WordNet's Top-Level: The OntoClean approach Proceedings of the Workshop OntoLex'2, Ontologies and Lexical Knowledge Bases, 2002

Guarino, N. & Welty, C. Evaluating ontological decisions with OntoClean Commun. ACM, ACM Press, 2002, 45, 61-65

Smith, B.; Williams, J. & Schulze-Kremer, S. The ontology of the gene ontology. AMIA Annu Symp Proc, Institute for Formal Ontology and Medical Information Science, University of Leipzig., 2003, 609-613.

Smith, B. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. J Biomed Inform, Department of Philosophy and National Center for Biomedical Ontology, University at Buffalo, Buffalo, NY 14260, USA. phismith@buffalo.edu, 2006, 39, 288-298

Obrst, Leo; Todd Hughes; Steve Ray. 2006. Prospects and Possibilities for Ontology Evaluation: The View from NCOR. Workshop on Evaluation of Ontologies for the Web (EON2006), Edinburgh, UK, May 22, 2006.

Obrst, Leo; Werner Ceusters; Inderjeet Mani; Steve Ray; Barry Smith. 2007 The Evaluation of Ontologies: Toward Improved Semantic Interoperability. Chapter in: Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, Christopher J. O. Baker and Kei-Hoi Cheung, Eds., Springer, 2007.

Gangemi, Aldo; Carola Catenacci1; Massimiliano Ciaramita1; Jos Lehmann; contributions by: Rosa Gil (in section 2.2). Francesco Bolici and Onofrio Strignano (in section 2.4). 2004. Ontology evaluation and validation: An integrated formal model for the quality diagnostic task.). OntoEval 2004.

Gangemi, A.; Catenacci, C.; Ciaramita, M.; Lehmann, J. 2005. A Theoretical Framework for Ontology Evaluation and Validation. In Proceedings of SWAP2005. http://www.loa-cnr.it/Papers/swap_final_v2.pdf

Gangemi, Aldo; Carola Catenacci; Massimiliano Ciaramita; and Jos Lehmann. 2006. Modelling ontology evaluation and validation. To appear in Proceedings of ESWC2006, Springer.

Lawrence Hunter, Mike Bada, K. Bretonnel Cohen, Helen Johnson, William Baumgartner, Jr. and Philip V. Ogren. "Ontology Quality Metrics," Center for Computational Pharmacology University of Colorado School of Medicine, October 8, 2007.

Lynette Hirschman , Jong C. Park , Junichi Tsujii , Limsoon Wong , and Cathy H. Wu. *Accomplishments and challenges in literature data mining for biology*. Bioinformatics 18: 1553-1561.

Proceedings of NIST 2007 Automatic Content Extraction Workshop (ACE), 2007. http://www.nist.gov/speech/tests/ace/ace07/.

Methods and Metrics for Ontology Evaluation Workshop (Sponsors: NIST and NIH), National Institute of Standards and Technology, Gaithersburg, MD, October 25-26.

Mani, I., B. Sondheim, D. House, L. Obrst. 1998. TIPSTER Text Summarization Evaluation: Final Report, MITRE technical report, Reston, VA, September, 1998.

Proceedings of NIST 2007 Automatic Content Extraction Workshop (ACE), 2007. http://www.nist.gov/speech/tests/ace/ace07/.

Background: Prior Technical Approaches

- Evaluation in use - Navigli et al. 2003, Porzel and Malaka 2005
  - **Best case:** Halo Project - Friedland et al. 2004
- Data-driven evaluation - essentially the fit between the ontology and a knowledge source e.g. a corpus - Brewster et al. 2004
- Gold Standard approaches, very common, used by for example, Cimiano et al. 2005
  - Major problem is the arbitrary choice of an ontology
  - Dellschaft and Staab 2006, proposed a method to derive IR/NLP type Precision/Recall/F-Measure

Background : Philosophical and Methodological Approaches

Methontology approach of Gomez-Perez:

- Focus on evaluating procedural or formative aspects ontology construction
- Criteria included: Consistency, Completeness, Conciseness, Expandability
- Some tools developed reflecting these approaches:
  - Lam et al. 2004, Knublauch et al. 2004, Alani 2005, 2006

OntoClean approach of Guarino and Welty:

- Philosophical approach based on theoretical principles:
- Proposed a set of "metaproperties":
  - Rigidity
  - Identity
  - Unity
- Much focus on "cleaning up" existing "ontologies" such as WordNet so as to make them more rigororous