Challenges in Commercializing Expert Knowledge Authoring

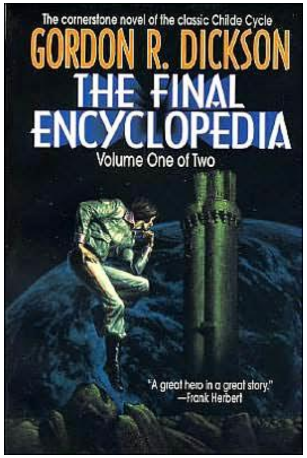
Vinay K. Chaudhri

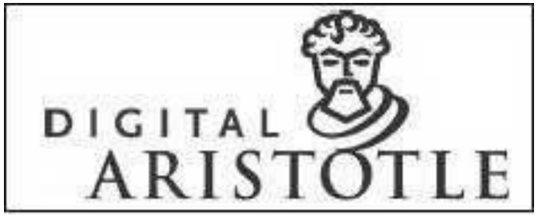
Acknowledgment

- AURA/Inquire Development Team
  - The original development work was funded by Vulcan Inc.
  - Eva Banik, Peter Clark, Roger Corman, Nikhil Dinesh, Debbie Frazier, Stijn Heymans, Sue Hinojoza, David Margolies, Adam Overholtzer, Aaron Spaulding, Ethan Stone, William Webb, Michael Wessel and Neil Yorke-Smith
  - Ashutosh Pande, Naveen Sharma, Rahul Katragadda, Umangi Oza
- Commercialization effort has been funded by SRI International

Vulcan’s Goals

- Build a ``Digital Aristotle’’ – a reasoning system capable of answering novel questions and solving advanced problems in a broad range of scientific disciplines





In 350 BC, Aristotle classified the world knowledge and introduced a system of logical reasoning

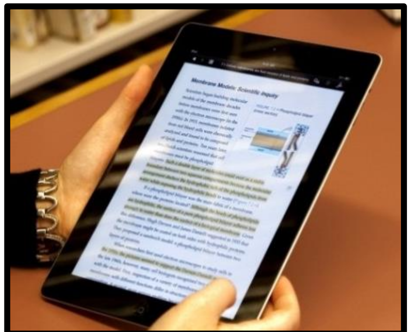
Realizing Digital Aristotle Vision

- Specific goals
  - Create knowledge representation for a textbook in a way that it can be used for answering questions and generating explanations
  - Create a platform technology that can be applied to multiple textbooks and multiple disciplines
- Promise: An ultimate digital tutor
  - Deep inquiry and dialog (e.g., follow up questions)
  - Precise student modeling (e.g., can pinpoint gaps in understanding)
  - Student engagement (e.g., as addictive as a game)

What we have achieved so far?

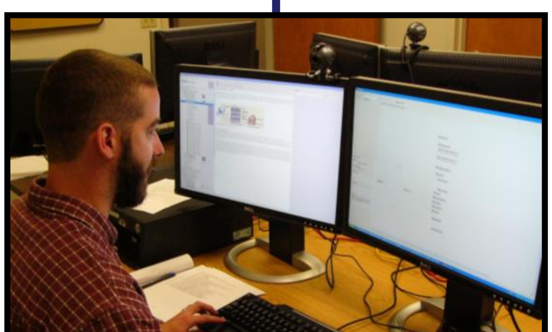
AURA

Embed Knowledge Representation in an Electronic Textbook



Find Real-World Use

2004 - 2009 | 2010 | 2011 | 2012-2013



AURA Authoring System
Physics, Chemistry, Biology

User Studies

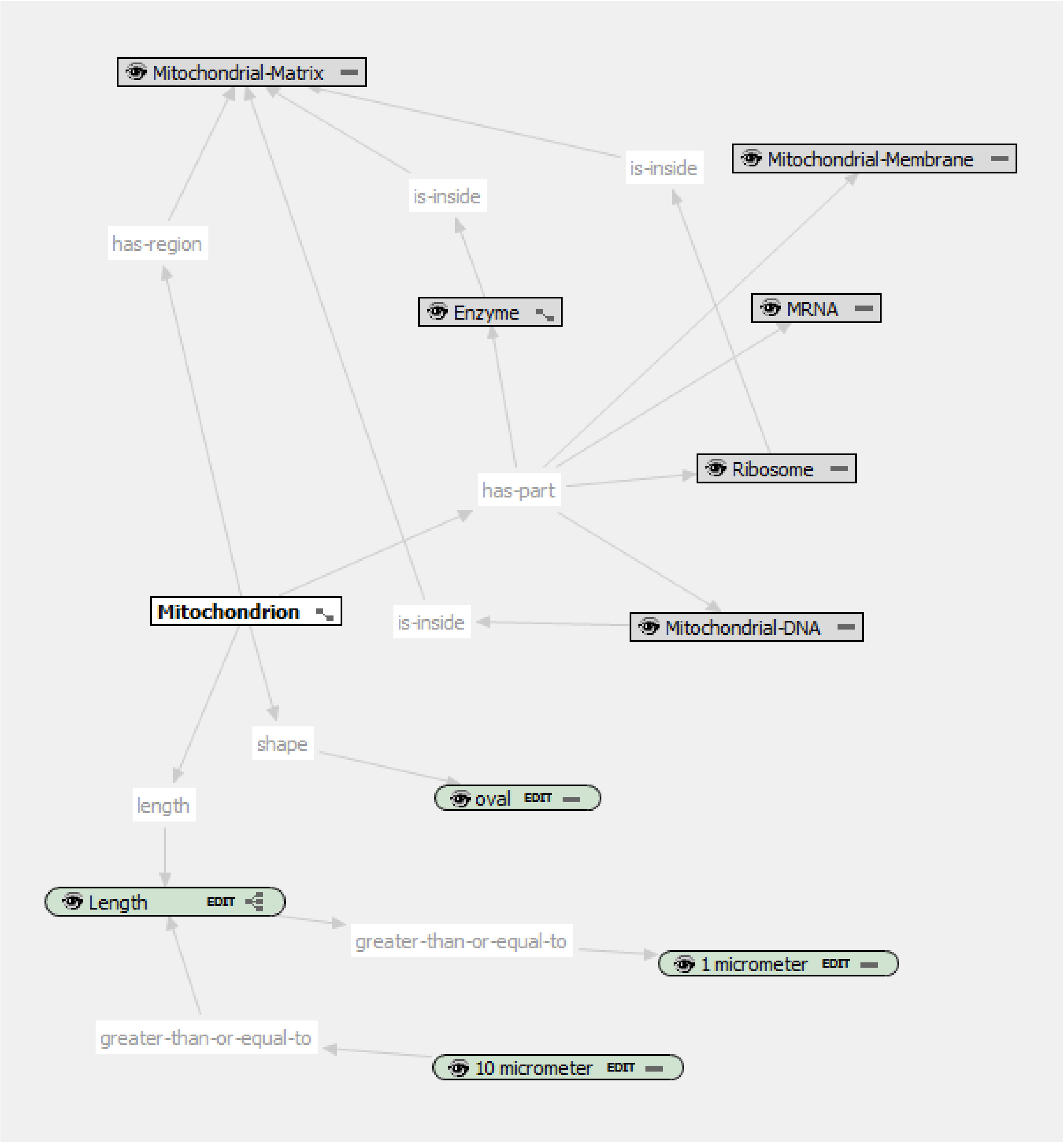
Outline

- Key differentiators in the technology
  - Knowledge authoring
  - Natural language Q/A
  - Natural language Generation
- Commercialization
  - Successes
  - Challenges

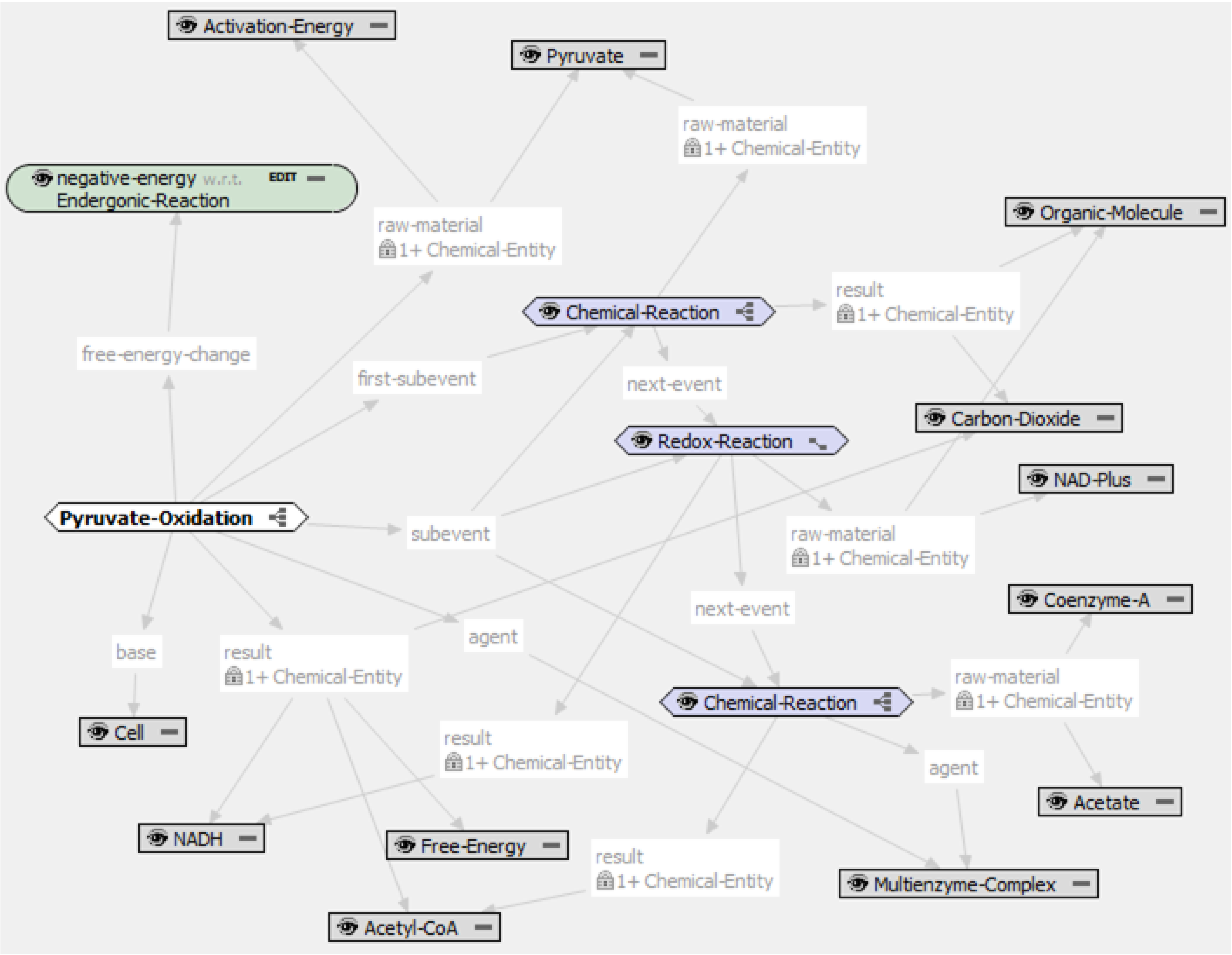
Knowledge Authoring in AURA

- Knowledge engineers provide a small library of domain independent representations
  - The Component Library (CLIB) contains classes representing physical actions, *e.g*., *Move, Attach, Penetrate*, and semantic relations, *e.g., agent, object, has-part (Barker, Clark, Porter, KCAP'01)*
    - See http://www.ai.sri.com/pub_list/864
- Biologists apply those representations to encode biology knowledge
  - AURA provides graphical editing
    - See http://www.ai.sri.com/pub_list/1545 and http://www.ai.sri.com/pub_list/865

Example Structure Representation



Formulated Knowledge



1) Determining Relevance and Pre-Planning

Pre-planning	Determining relevance, Diagram analysis, Pre-planning Status Labeling: Relevant, Irrelevant (closed)

2) Reaching Consensus

Universal Truth authoring, Concept chosen	QA check

3) Encoding Planning

Group common UTs, Identify KR/KE issues, Identify already encoded, Write how to encode	Planning, QA check Status Labeling: Encoding Complete, KR Issue (closed)

4) Encoding

Encode, File KR JIRA issues	QA check Status Labeling: Encoding Complete, KE Issue (closed)

5) Key Term Review

KR evaluated by modeling expert and SME, Encoder makes changes	KR evaluated by modeling expert and SME QA check

6) Question-Based Testing

Use Minimal Test Suite, File reasoning JIRA issues, Encoder fills KB gaps	QA check with screenshots of 'Passing' comparison and relationship questions

KB_Bio_101 Statistics

AURA

Regarding Class Axioms:

# Classes	# Relations	# Constants	Avg. # Skolems / Class	Avg. # Atoms / Necessary Condition	Avg. # Atoms / Sufficient Condition
6430	455	634	24	64	4

# Constant Typings	# Taxonomical Axioms	# Disjointness Axioms	# Equality Assertions	# Qualified Number Restrictions
714	6993	18616	108755	936

Regarding Relation Axioms:

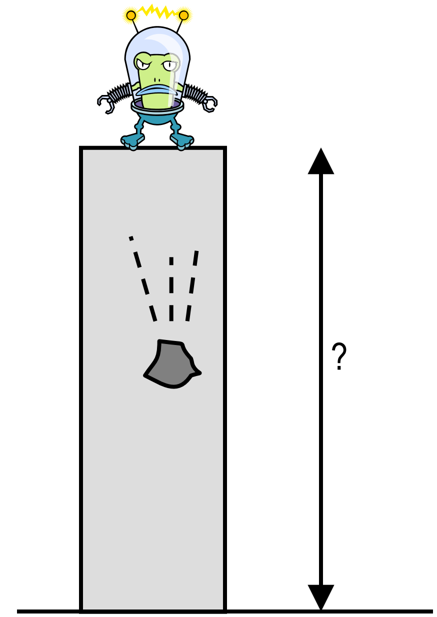
# DRAs	# RRAs	# RHAs	# QRHAs	# IRAs	# 12NAs / # N21As	# TRANS + # GTRANS
449	447	13	39	212	10 / 132	431

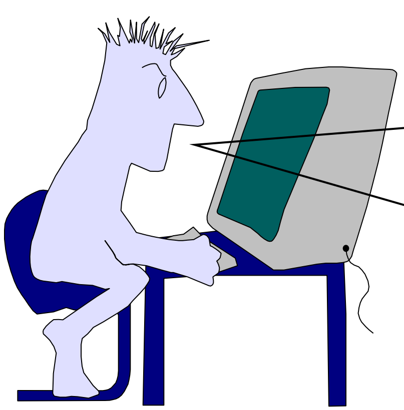
Regarding Other Aspects:

# Cyclical Classes	# Cycles	Avg. Cycle Length	# Skolem Functions
1008	8604	41	73815

Example of Question Formulation

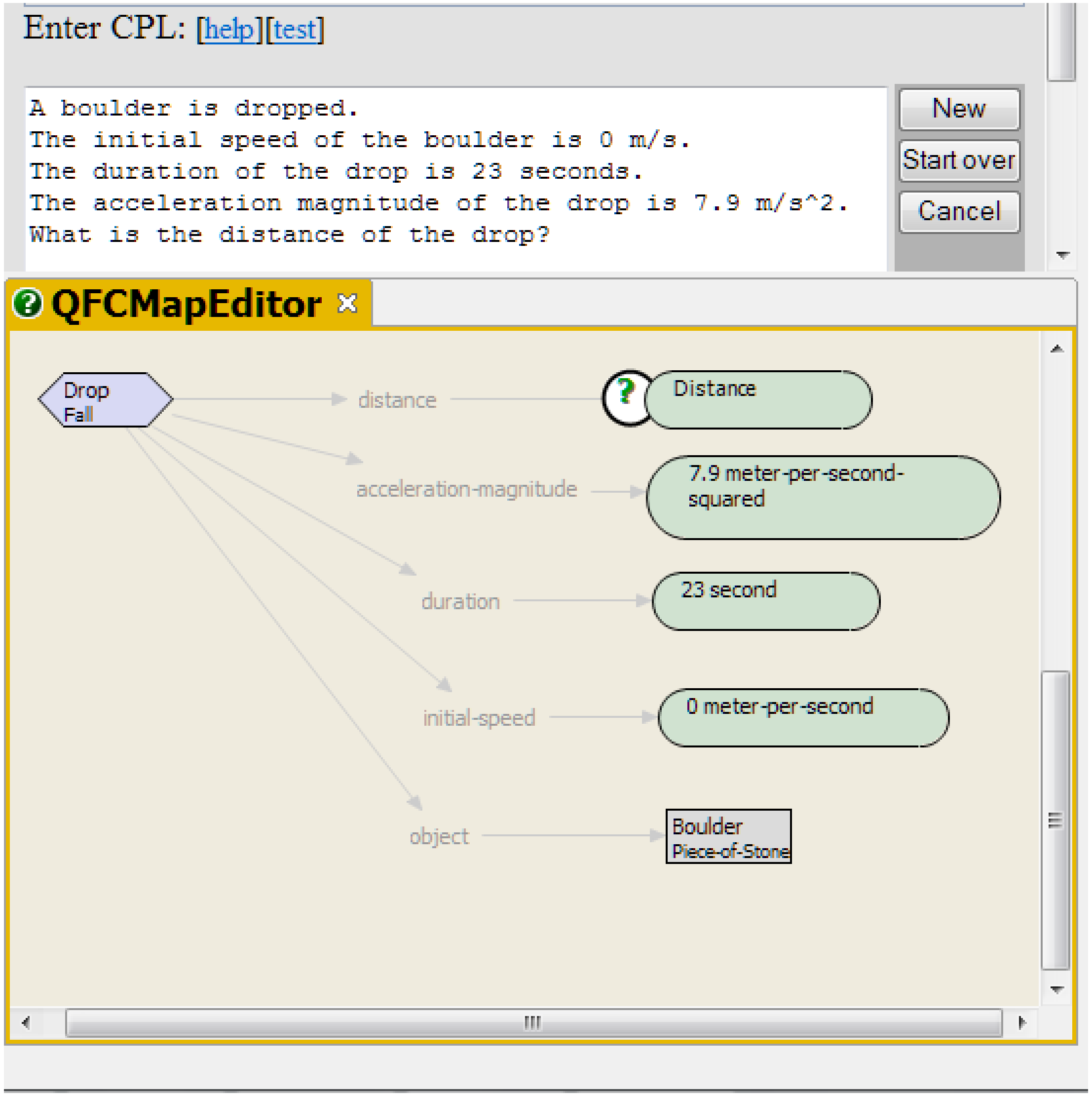
An alien measures the height of a cliff by dropping a boulder from rest and measuring the time it takes to hit the ground below. The boulder fell for 23 seconds on a planet with an acceleration of gravity of 7.9 m/s^2. Assuming constant acceleration and ignoring air resistance, how high was the cliff?





A boulder is dropped.
The initial speed of the boulder is 0 m/s.
The duration of the drop is 23 seconds.
The acceleration of the drop is 7.9 m/s^2.
What is the distance of the drop?

Example Feedback from the System



Lookup

1. What are the types of X?
2. What is the structure of X?
3. What are the steps of X?
4. What is/are the slotA of a X?

Identify

1. Given a set of properties of X, what is an X an instance of?

Compare

1. What are the differences/similarities between X and Y?
2. What are the functional differences/similarities between X and Y?
3. What are the structural differences/similarities between X and Y?
4. What is the energetic difference between X and Y?
5. What are the differences/similarities between the SlotA of X and the SlotA of Y?
6. What are the differences/similarities between the ConceptA slotB of X and the ConceptB slotB of Y?

Relate

1. What is the relationship between X and Y?
2. What is the qualitative relationship between X and Y?
3. What is the qualitative relationship between PropertyA of X and PropertyB of Y?
4. What is the qualitative relationship between PropertyA of X and the function of Y?
5. What is the energetic relationship between X and Y?
6. X is to Y as Z is to what?

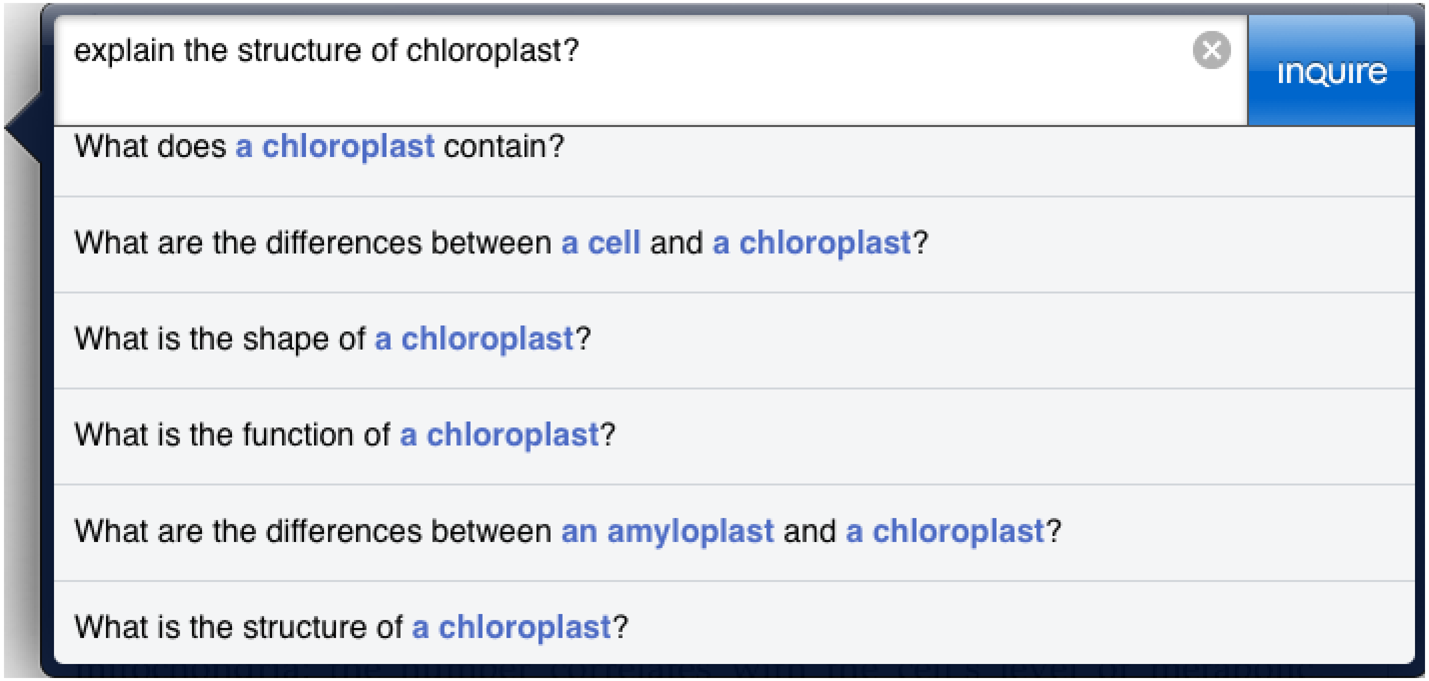
Describe

What is X?

Determine

1. How many Y are SlotA of a X?
2. Is it true that X is a Y?
3. [In X], what acts as Y [in Z]?
4. What structures of X facilitate Y?
5. What structures of X facilitate the function of X?
6. If A is removed from B, what events will be affected?
7. If A is removed from B, will C be affected?
8. Regulation and Energy Flow questions (20)

Suggesting Questions



Natural Language Generation

AURA

What is pyruvate oxidation?

Definition of a pyruvate oxidation

A step in the process of cellular respiration where pyruvate, formed during glycolysis, (A) enters the mitochondria and is oxidized to a compound called acetyl CoA which enters the citric acid cycle.

Pyruvate oxidation is a type of: catabolic pathway. (B)

Properties of a pyruvate oxidation

Free energy change (C)

negative with respect to an endergonic reaction

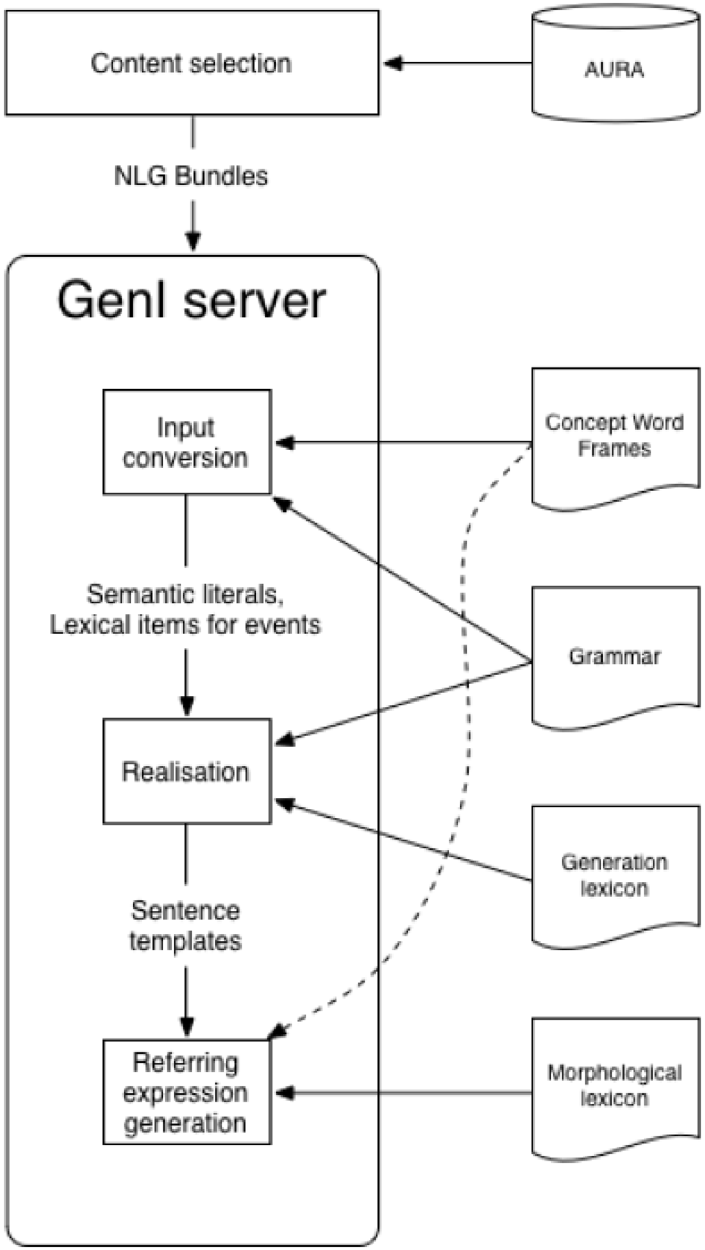
Participants of a pyruvate oxidation

pyruvate is converted by a multienzyme complex in a cell to an acetyl CoA, carbon dioxide and an NADH. This process transforms activation energy to free-energy. (D)

Steps of pyruvate oxidation (E)

1. Chemical reaction consumes pyruvate and produces carbon dioxide and an organic molecule
2. Redox reaction consumes NAD plus and an organic molecule and produces an NADH
3. Chemical reaction — a multienzyme complex converts acetate and coenzyme A to an acetyl CoA

NLG Architecture



Outline

- Key differentiators in the technology
  - Knowledge authoring
  - Natural language Q/A
  - Natural language Generation
- Commercialization
  - Successes
  - Challenges

Commercialization Challenges

AURA

- This innovation is too long-term and cannot be immediately translated into profits
- Publishers are too daunted by KB authoring, and instead, we need to engage the textbook authors
  - Show the value of using conceptual representation in improving a discipline
- Further research is needed (at the intersection of AI and education)
- Product-focused R&D is required
- Find sponsors who are not driven by short-term gains (e.g., foundations)

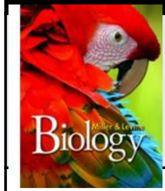
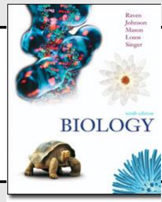
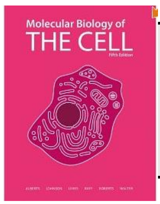
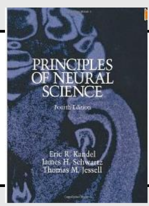
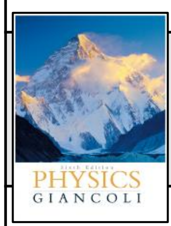
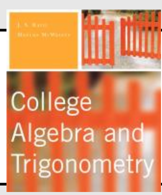
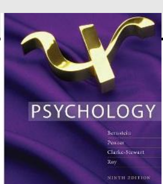
Challenge 1: Long-term innovation

- Ontology-based question answering is too radical a change for high school education
  - Q/A is not a common place technology even for bio-informatics researchers
  - Education innovations usually begin at graduate level and trickle down to lower grade levels

Challenge 2: Publishers too daunted

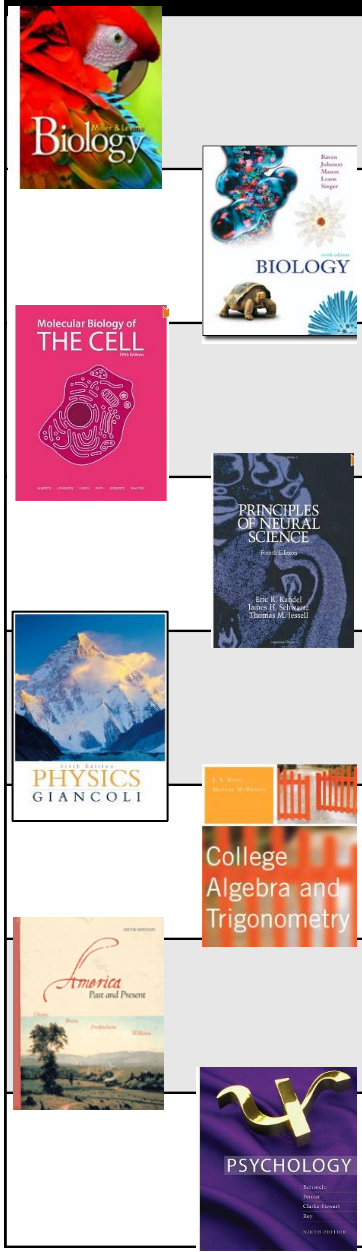
- Publishers are driven by immediate profits
  - They need fully automated technology that can be applied to lots and lots of books
- Need to appeal to textbook authors
  - Model creation needs to become an integral part of textbook authoring
  - Just like we manually build figures, we could manually build conceptual models
    - These models are then available to an electronic textbook for reasoning and question answering

Generalization to multiple textbooks

Textbook	
	Middle school biology
	Comparable to Campbell biology
	Cell biology
	Neuroscience
	Introductory college physics
	Introductory college algebra
	Introductory college US history
	Introductory college psychology

Generalization to multiple textbooks

Textbook



General Aspects:

1. Conceptual and qualitative knowledge cuts across domains
2. Some domains are more mathematical than others and require mathematical/symbolic problem solving
3. Challenges in representing Campbell also exist in other disciplines: models, hypotheses, experiments

Unique aspects:

1. Each domain requires domain-specific vocabulary design
2. Each domain has some new question formulation challenges
3. Each domain has some new unique representations needs

Challenge 3: Further research

- We do not have ontology designs for capturing all of textbook knowledge
  - For example, see our FOIS paper on content modeling challenges
  - We can currently model only 40-50% of textbook knowledge
  - We need sustained ontology research to capture greater fractions of textbook knowledge

Challenge 4: Product-focused R&D

- How much of the textbook do we actually need to capture?
  - What is the minimal viable representation?
  - How much of the representation can be incrementally added?
- Should the answer be limited to just the chapter studied?

Challenge 5

- Need non-profit driven funding
  - Academic research sources
  - Foundation and philanthropic support

Next Steps

- Continue to leverage on the successes
- Identify and work with Foundation sponsors

Thank You!