Knowledge Formulation, Reasoning and Question Answering in Project Halo

Mark Greaves
Vulcan Inc.
April 13, 2010
Outline

- **Project Halo Overview**
  - Textbooks You Can Talk To, An Analytic Encyclopedia, and Tractatable KRs for Defaults

- **Deeper Dive: Crowdsourcing Data**
  - Extending Wikis: Vulcan’s Semantic MediaWiki+
  - A Gallery of Semantic MediaWiki+ Applications

- **Ultrapedia: An Analytic Encyclopedia**
  - Ultrapedia Under the Hood
  - Ultrapedia Demo (Screenshots in Appendix)
  - The Future
Outline

- Project Halo Overview
  - Textbooks You Can Talk To, An Analytic Encyclopedia, and Tractatable KRs for Defaults

- Deeper Dive: Crowdsourcing Data
  - Extending Wikis: Vulcan’s Semantic MediaWiki+
  - A Gallery of Semantic MediaWiki+ Applications

- Ultrapedia: An Analytic Encyclopedia
  - Ultrapedia Under the Hood
  - Ultrapedia Demo (Screenshots in Appendix)
  - The Future
Envisioning the Digital Aristotle for Scientific Knowledge

- Inspired by Dickson’s Final Encyclopedia, the HAL-9000, and the broad SF vision of computing
  - The “Big AI” Vision of computers that work with people

- The volume of scientific knowledge has outpaced our ability to manage it
  - This volume is too great for researchers in a given domain to keep abreast of all the developments
  - Research results may have cross-domain implications that are not apparent due to terminology and knowledge volume

- “Shallow” information retrieval and keyword indexing systems are not well suited to scientific knowledge management because they cannot reason about the subject matter
  - Example: “What are the reaction products if metallic copper is heated strongly with concentrated sulfuric acid?” (Answer: Cu$^{2+}$, SO$_2$(g), and H$_2$O)

- Response to a query should supply the answer (possibly coupled with conceptual navigation) rather than simply list 1000s of possibly relevant documents
How do we get to the Digital Aristotle?

- **What the Digital Aristotle requires:**
  - Technology to enable a global, widely-authored, very large knowledge base (VLKB) about human affairs and science,
  - Technology that answers questions and proactively supplies information,
  - Technology that uses powerful reasoning about rules and processes, and
  - Technology that can be customized in its content and actions for individual organizations or people

- **Vulcan’s Goals**
  - Address the problems of **scale** and **brittleness** in Knowledge Bases
    - Incorporate large numbers of SMEs in KB construction and maintenance
    - Actively support knitting KBs together
  - Have high **impact**
    - Show that the Digital Aristotle is possible in a specific area
    - Change our experience of AI
    - Have quantifiable, explainable metrics
  - Be a **commercializable** approach

- **Project Halo** is a concrete research program that addresses these goals
In 2004, Vulcan funded a six-month effort to determine the state-of-the-art in fielded “deep reasoning” systems

- Can these systems support reasoning in scientific domains?
- Can they answer novel questions?

Three teams were selected: SRI, Cycorp, and Ontoprise GmbH

Evaluation

- Teams were given 4 months to formulate the knowledge in 70 pages from the US AP (Advanced Placement) Chemistry syllabus
- Systems were sequestered and run by Vulcan against 100 novel AP-style questions (hand coded queries)
- Exams were graded by chemistry professors using AP methodology

Best scoring system achieved roughly an AP3 (on our very restricted syllabus)

Full Details in *AI Magazine* 25:4, “Project Halo: Towards a Digital Aristotle”, and at www.projecthalo.com
From the Halo Pilot to the Halo Project

- Halo Pilot Results
  - Much better than expected results on a very tough evaluation
  - Most failures attributed to modeling errors due to contractors’ lack of domain knowledge
  - Expensive: O($10,000) per page, per team

- Project Halo Goal: To determine whether tools can be built to facilitate robust knowledge formulation, query and evaluation by domain experts, with ever-decreasing reliance on knowledge engineers
  - Can SMEs build robust question-answering systems that demonstrate excellent coverage of a given syllabus and the ability to answer novel questions
  - Will SMEs be capable of posing questions and complex problems to these systems?
  - Do these systems address key failure, scalability and cost issues encountered in the Pilot?

- Scope: Selected portions of the AP syllabi for chemistry, biology and physics

- Two competing teams/approaches (F-Logic, Concept Maps/KM)

- Evaluation and downselect in September 2006
Motivation
- “Read a Chapter of a text and answer questions at the back of the chapter” (Three Open Problems in AI, JACM’03)
- “Build a Knowledge Base by Reading a Textbook” (Some Challenges for Computational Intelligence, JACM’03)

Focus on fundamental hard sciences where knowledge is explicitly written down
- Physics, Chemistry, and Biology
- Mid-level difficulty: 1st-year college course

Choose a widely accepted test for competence
- Advanced Placement Test
- The AP test is just a metric. The system capability should be general enough to answer a broader set of questions

Start with a manageable scope and work outwards
- ~100 pages of AP syllabus in Physics, Chemistry, Biology

Concept of Operations

Domain Experts Enter Knowledge
Domain Experts add knowledge to the AURA Knowledge Base and import knowledge using the mapping tool

Users ask questions and get answers and explanations

AURA Answers Questions
Based on These Results, Team SRI was Selected over Team Ontoprise
The Halo Project Today – Three Efforts in Systems AI

- AP-level Knowledge Entry and Question Answering Technology (HaloBook)
- SME entry and use of defaults and rule knowledge (SILK)
- Scaling up Participation (Semantic Wikis / Ultrapedia)
- Aura is a tool to help users formalize knowledge.
- Aura can then reason with that knowledge.
- So users can ask questions and understand the answers.
Knowledge Formulation in AURA

The left screen displays the AURA progress manager, the source textbook from which the users select the passages for knowledge formulation.

The right screen displays the palette, concept details, and the AURA concept map that allows the user to graphically formulate knowledge.

AURA concept map with equations.

AURA converts Concept Maps, equations, and tables into declarative knowledge to support question answering.
Users formulate questions in a restricted English dialect (CPL)

The users can graphically edit the system’s interpretations

The system reasons with the interpretation to answer questions
AURA Phase II Evaluation (2008)

EXPERIMENTAL PROCEDURE

- Domain experts used AURA to author KBs for 50 page sections of textbooks in Biology, Chemistry, and Physics (using a set of reference questions to test their KBs)
- The KBs were tested against a set of novel questions, independently generated by AP teachers of those domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Known Question Set</th>
<th>Novel Question Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>71%</td>
<td>47%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>73%</td>
<td>24%</td>
</tr>
<tr>
<td>Physics</td>
<td>79%</td>
<td>36%</td>
</tr>
</tbody>
</table>

EXPECTED RESULTS

- Domain Experts created high quality KBs in all domains
- We had good results in Biology giving us confidence to propose encoding a full book as a good next scaling target

UNEXPECTED RESULTS

- We had significant challenges in Chemistry and Physics with novel questions, and we identified areas for system improvements
Scaling Up to a Textbook You Can Talk To

- Offshore experiment in KB construction
  - India = educated population, native English
  - Aura tested at IJCAI 2008 with IIIT-Hyderabad students
  - Partner with Indian companies
    - Does the lower cost of offshore labor offset higher management costs?
    - Compare two collaborative knowledge entry designs to find the best
  - Pilot with the implementation phase syllabus in Biology; collaborative knowledge entry design

- With Educator/Publisher partners, design a Question-Answering Biology Textbook

- Murray Burke’s Challenge: A Calculator for Biology
  - Enhanced iPad-based textbook reader with AURA-based question answering
  - Questions are answered with reference to the specific knowledge in the text
  - A brand-new capability with huge educational potential
Offshore Knowledge Factory Experiment (2009)

EXPERIMENTAL PROCEDURE

- Two Indian firms were given the same Biology knowledge formulation problem used in the AURA Phase II Evaluation
- Both firms were asked to use three-person domain teams to collaboratively enter the knowledge

<table>
<thead>
<tr>
<th>Domain Team</th>
<th>Known Question Set</th>
<th>Novel Question Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company A</td>
<td>85%</td>
<td>75%</td>
</tr>
<tr>
<td>Company B</td>
<td>75%</td>
<td>68%</td>
</tr>
</tbody>
</table>

EXPECTED RESULTS

- Offshore teams were able to replicate the KB creation process
- KFEs were able to successfully collaborate

UNEXPECTED RESULTS

- We were expecting the scores in the range of 50%
- Breadth and depth of syllabus coverage is superior
HaloBook is a concept for a new kind of electronic textbook that contains an underlying knowledge-base of the book’s contents that answers the reader’s questions and provides tailored instruction.

This briefing presents a sketch of how a HaloBook might look in 2015 assuming continued Halo development and modest success in solving research challenges.

An initial HaloBook prototype will be developed in 2010.
A HaloBook user can:

- **READ** the contents dynamically, interactively
- **ASK** questions and get explanations on any subject in the book
- **LEARN** to master the subject through individualized tutoring
- **CREATE** and explore their own conceptualizations of the material

**HaloBook Functions**

**The mitotic phase alternates with interphase in the cell cycle: an overview**

Mitosis is just one part of the cell cycle (FIGURE 12.4). In fact, the **mitotic (M) phase**, which includes both mitosis and cytokinesis, is usually the shortest part of the cell cycle. Mitotic cell division alternates with a much longer interphase, which often accounts for about 90% of the cycle. It is during interphase that the cell grows and copies its chromosomes in preparation for cell division.

Interphase can be divided into subphases: the **G₁ phase** ("first gap"), the S phase, and the **G₂ phase** ("second gap"). During all three subphases, the cell grows by producing proteins and cytoplasmic organelles. However, chromosomes are duplicated only during the S phase.

**FIG 12-4.** The cell cycle. In a dividing cell, the mitotic (M) phase alternates with interphase, a growth period.
A HaloBook user can:

- Read the material, page by page, as written
- View tailored presentations on requested topics
- Preview sections through a dynamically generated skim of the section’s main concepts
- Conduct all common eBook activities (lookup definitions, follow links to related material, etc.)
HaloBook Function: 

ASK

A HaloBook user can:

- Ask questions, in simplified natural language
- Get explanations, tailored to the student’s preferences and understanding
- Get explanations, as dynamically generated graphics, animations, and simulations
HaloBook Function: ASK

A HaloBook user can:

• Ask questions, in simplified natural language

• Get explanations, tailored to the student’s preferences and understanding

• Get explanations, as dynamically generated graphics, animations, and simulations
Rules Knowledge in Project Halo

- **Address key brittleness issues in Aura (and in AI in general)**
  - Default, hypothetical, action/event, and process knowledge
  - Integration of external knowledge bases including Semantic Web

- **SILK: Suite of core knowledge representation and reasoning (KR) modules**
  - Provide defaults, hypotheticals, actions, and processes capabilities
    - First Focus: Combine defaults with as much as possible of other established features for monotonic (DB, classical, ontology). Default flavor pervades the KR.
    - Second Focus: Hypotheticals/Actions/Processes. Key ideas: advanced defaults and rules
  - Design for distributed algorithms and platform for high scalability
    - Focus: Incremental update/merge, with distributed dynamic import
    - Key ideas: dependency analysis, precomputation
  - Progressively/iteratively extend with new expressive features and algorithms
  - Interoperate via KR and SOA standards with other systems/sources, including web sources

- **Integration of the above**
  - Into Aura, to significantly boost AP performance
  - Into SMW+ or other wiki/Web2.0 environment, for knowledge acquisition
  - As a stand-alone KR technology
## SILK V1 Reasoning Engine

### First prototype of SILK’s Hyper Logic Programs KR capability
- Implemented using the Flora Logic Programming system

### Semantic Rule System Capabilities

<table>
<thead>
<tr>
<th></th>
<th>Jena</th>
<th>Flora</th>
<th>SILK V1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simple rules</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal logical connectives and functions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grandfather(?x,?g) :- parent(?x,?y) and father(?y,?g).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Higher-order</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicates or formulas can be variables in other formulas</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>?Propn :- says(?Src,?Propn) and topic(?Propn,?T) and trusted(?Src,?T).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Motivating Requirements:</em> meta-reasoning, KR macros, KB translation &amp; integration, reasoning control, provenance, navigation in KA, multi-agent &amp; nested belief, context, modals. Plus – the Web is about meta-data.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Defaults</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rules can override each other or have exceptions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{g1} gravity(?Q,?G) :- APproblem(?Q) and gravity(Earth,?G).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{g2} gravity(Q22,?G) :- gravity(Mars,?G).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overrides(g2,g1). APproblem(Q22). gravity(Earth,9.8N). gravity(Mars,3.7N).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Motivating Requirements:</em> robust KB merging/updating, process causality, commonsense, natural language &amp; KA, inductive/scientific learning, conflict, argumentation, OO inheritance, policy &amp; regulation/law, import of classical knowledge</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Scaling and Performance

- Computational scalability to 1M+ Rules

<table>
<thead>
<tr>
<th></th>
<th>Jena</th>
<th>Flora</th>
<th>SILK V1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Defaults</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rules can override each other or have exceptions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{g1} gravity(?Q,?G) :- APproblem(?Q) and gravity(Earth,?G).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{g2} gravity(Q22,?G) :- gravity(Mars,?G).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overrides(g2,g1). APproblem(Q22). gravity(Earth,9.8N). gravity(Mars,3.7N).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Motivating Requirements:</em> robust KB merging/updating, process causality, commonsense, natural language &amp; KA, inductive/scientific learning, conflict, argumentation, OO inheritance, policy &amp; regulation/law, import of classical knowledge</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Green** = yes, i.e., good capability
- **Yellow** = partial, i.e., along the road but not there yet
## Feature Comparison – SILK vs. other rule systems

<table>
<thead>
<tr>
<th>Level</th>
<th>Groups of Features</th>
<th>SILK2</th>
<th>Flora</th>
<th>RIF-BLD</th>
<th>Jena</th>
<th>Onto-broker</th>
<th>Jess</th>
<th>IBM C.R.</th>
<th>DLV</th>
<th>SQL</th>
<th>SPA-RQL</th>
<th>Common Logic</th>
<th>OWL2 RL</th>
<th>OWL2 DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>Horn chain. etc.</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>R.</td>
<td>Y</td>
<td>R.</td>
<td>R.</td>
</tr>
<tr>
<td>Advanced</td>
<td>(Level summary)</td>
<td>Most!</td>
<td>lots</td>
<td>some</td>
<td>some</td>
<td>some</td>
<td>some</td>
<td>some</td>
<td>some</td>
<td>some</td>
<td>some</td>
<td>some</td>
<td>some</td>
<td>some</td>
</tr>
<tr>
<td>Functions</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Closed-World</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>R.</td>
<td>R.</td>
<td>most</td>
<td>R.</td>
<td>R.</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Higher-Order</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>R.</td>
<td>N</td>
<td>N</td>
<td>R.</td>
<td>R.</td>
<td>Y</td>
<td>R.</td>
<td>bit</td>
</tr>
<tr>
<td>Actions</td>
<td></td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Base Defaults</td>
<td></td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Webized</td>
<td></td>
<td>Y</td>
<td>R.</td>
<td>Y</td>
<td>Y</td>
<td>R.</td>
<td>R.</td>
<td>R.</td>
<td>R.</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Hyper</td>
<td>(Level summary)</td>
<td>1st!</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>H-O. Defaults</td>
<td></td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Weak. Classi.</td>
<td></td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Misc.</td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Other Expres.</td>
<td></td>
<td>Dev.</td>
<td>inherit.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>events</td>
<td>-</td>
<td>disju.</td>
<td>R.</td>
<td>R.</td>
<td>classical</td>
<td>-</td>
<td>classic.</td>
</tr>
<tr>
<td>Efficiency</td>
<td></td>
<td>good</td>
<td>good</td>
<td>NA</td>
<td>fair</td>
<td>good</td>
<td>fair</td>
<td>good</td>
<td>poor</td>
<td>good</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Dev. = Developing; R. = Restricted; C.R. = Common Rules; disju. = disjunctive
Outline

- Project Halo Overview
  - Textbooks You Can Talk To, An Analytic Encyclopedia, and Tracatable KRs for Defaults

- Deeper Dive: Crowdsourcing Data
  - Extending Wikis: Vulcan’s Semantic MediaWiki+
  - A Gallery of Semantic MediaWiki+ Applications

- Ultrapedia: An Analytic Encyclopedia
  - Ultrapedia Under the Hood
  - Ultrapedia Demo (Screenshots in Appendix)
  - The Future
Project Halo and the Argument for Wikis

- **Project Halo is a staged R&D effort to build very large, useful question-answering systems**
  - Initially in natural science, to answer questions at the AP level
  - Traditional search systems (Google/Yahoo/Microsoft, Endeca) do not work for this task

- **Three Project Halo challenges:**
  - Knowledge Formulation: Can you build the knowledge bases?
  - Question Formulation: Can you query the knowledge bases?
  - Question Answering and Explanation Generation: Can you get the answers?

- **Knowledge Formulation Challenge: create technology to build very large, computer-processable knowledge bases**
  - Millions of interlinked assertions, rules, and patterns, structured to support for question-answering algorithms, built and maintained in a cost-effective, reliable way
  - Require solutions to certain AI issues that limited previous attempts at large-scale KBs

- **Wikis are one way to crowdsource logically simple information**
  - Highly reliable, Internet-scale, and incredibly cheap
  - We have developed a *semantic wiki* which adds data and schema authoring to text wikis
Wikipedia is Fantastic, but...

- **Most Wikipedia data cannot be easily exploited by computers**
  - Searching is limited to keywords only, and the result unit is the whole article
  - No easy way to extract from Wikipedia, for example:
    - Sci-Fi movies after Y2K that cost < $10M while gross > $30M
    - All Porsche models that accelerate less than 6 seconds
    - Skyscrapers in China higher than 50 stories, built before 2000
    - Soccer player with jersey #11 from a club with a home stadium with more than 40000 seats, who were born in a country with more than 10M people

- **Tables in articles are (mostly) manually built**
  - By definition out of date, often inconsistent with text

- **Category hierarchy is inconsistent and arbitrary**

**Enter Semantic Wikis!**

Blend the ease and *consensus* of a wiki with the *query power* of a database
Semantic Wikis in One Slide

- **Semantic Wikis apply the wiki concept to data**
  - Authoring includes instances, data types, vocabularies, classes
  - Natural language text used for explanations
  - Automatic list generation from structured data, basic analytics, database imports
  - Reuse of wiki knowledge

- **Each wiki page describes a concept, which can have associated properties and values in addition to the text**
  - In classic SMW, this is manually encoded on the page by an article author
    - Just like regular wiki markup for sections, links, etc.
    - Semantic markup looks like `[property::value]`
  - Properties can be introduced whenever needed
    - “Seattle has population 602,000”
    - “Seattle is in Washington”

- **Semantic can also come from databases and natural language extraction systems**
Vulcan’s Semantic MediaWiki+ In Use

- **SMW+ is open-source (GPL)**
  - ~16,000 downloads; stable core since Sep 08
  - $10^6$-$10^7$ pages under SMW+ control, in wikis we know
  - Can be purchased, with support, from Ontoprise for €200.
  - Regular users’ group meetings (next SMWCon 5/23-24 in Boston, also European meetings on Sept 18-19)
  - A growing developer community at [http://smwforum.ontoprise.com](http://smwforum.ontoprise.com)

- **Example Commercial Users**
  - UNESCO: Ocean Teacher Encyclopedia
  - Siemens: Corporate library management
  - Volkswagen: Terminology management
  - PriceWaterhouse Coopers: accounting collaboration
  - Mayo Clinic/NIH: HL7 and BioMedGT vocabularies
  - LeveragePoint: strategic price consulting
Example: Skyscrapers in China higher than 50 stories, built before 2000

ASK/SPARQL query target

```{{#ask:
[[Category:Skyscrapers]]
[[Located in::China]]
[[Floor count::>50]]
[[Year built::<2000]]
...}}
Outline

- **Project Halo Overview**
  - Textbooks You Can Talk To, An Analytic Encyclopedia, and Tractatable KRs for Defaults

- **Deeper Dive: Crowdsourcing Data**
  - Extending Wikis: Vulcan’s Semantic MediaWiki+
  - A Gallery of Semantic MediaWiki+ Applications

- **Ultrapedia: An Analytic Encyclopedia**
  - Ultrapedia Under the Hood
  - Ultrapedia Demo (Screenshots in Appendix)
  - The Future
- Automatically populate tables
- Just the data you want, at the level you want
- Calendars and timelines
- Workflows
- Custom Reports
- Form-oriented inputs
- Notifications via email/RSS
- MS Office integration
From Excel-based patent tracking to flexible Wiki-based data entry and query tool
Collaborative Proposal Management at BT with SMW+

Active Bid Viewer
Service Desk Selector
Edit Rsnum: Rs1815739

Warning: You are not logged in. Your IP address will be recorded in this page's edit history.

SNP: sprinting vs endurance muscles

Summary

<table>
<thead>
<tr>
<th>Geno</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C;C)</td>
<td>possibly increased sprint/power performance</td>
</tr>
<tr>
<td>(C;T)</td>
<td>mix of sprinting &amp; endurance muscles</td>
</tr>
<tr>
<td>(T;T)</td>
<td>possibly increased endurance</td>
</tr>
</tbody>
</table>

Free text:

This [[SNP]], in the [[ACTN3]] gene, encodes a premature stop codon in a protein likely to be important in muscle function. The polymorphism alters position 577 of the alpha-actin-in-3 protein, and while the most common nucleotide at this position, (C), encodes an arginine (amino acid code R), the stop codon is referred to as X. Hence, the SNP is referred to as R577X, with homozygotes being either RR or XX and heterozygotes being RX.

The main report studying a relatively small number of Australian elite (ie -Olympic) athletes found that, at least in females, the R allele (ie [[rs1815739]](C)) is associated with sprinters, while the X allele...
**Category:** BGT Beckers Nevus (B4097)

**BGT_Beckers_Nevus(B4097)**

**Lexical**

- **Concept Code:** B4097  
- **Preferred Name:** Becker's Nevus  
- **Coding Scheme:** BGT (01.01)  
- **Synonym:** Becker's Nevus (PT) (Source: NCI) / Linear Papular Ectodermal Mesodermal Hamartoma (SY) (Source: NCI) / Pigmented Hairy Nevus of Becker (SY) (Source: NCI) / Progressive Cribriform and Zosteriform Hyperpigmentation (SY) (Source: NCI) / Pigmented Hairy Epidermal Nevus (SY) (Source: NCI) / Melanosis Neviformis (SY) (Source: NCI)  
- **URI:** urn:oid:2.16.840.1.113883.3.26.1.2.B4097

**Properties**

- **NCI/META/CUI:** CL103247  
- **Semantic Type:** Disease or Syndrome

**Associations**

- **Parent:** Hamartoma *(primitive)*  
- **Parent:** Non_Neoplastic_Nevus *(primitive)*  
  
  Every instance of **Beckers_Nevus** Disease_Has_Primary_Anatomic_SITE only instances of **Skin** if it Disease_Has_Primary_Anatomic_SITE anything at all.  
  Every instance of **Beckers_Nevus** Disease_Has_Associated_Anatomic_SITE only instances of **Integumentary_System** if it Disease_Has_Associated_Anatomic_SITE anything at all.  
  Every instance of **Beckers_Nevus** Disease_Has_Finding only instances of **Cutaneous_Involvement** if it Disease_Has_Finding anything at all.
Video Search: Semantic Wikis for A New Problem

- Social tag-based characterization
- Keyword search over tag data
- Inconsistent semantics
- Easy to engineer

- Algorithm-based object characterization
- Database-style search
- Consistent semantics
- Extremely difficult to engineer

Semantics Wikis for Video Search

- Social database-style characterization
- Database search + wiki text search
- Semantic consistency via wiki mechanisms
- Easy to engineer

Demo on Seahawks Video
Seattle Seahawks Game Video Indexing Wiki

- Commercial Look/Feel
- Play-by-play video search
- Highlight reel generation
- Search on crowd-defined patterns ("touchdowns with big hits")
- Tree-based navigation widget
- Very favorable economics
SMW User Forum Wiki

SMW+ User Manual

User manual article for SMW+ 1.4.6

By: Daniil

SMW+ is a semantic enterprise wiki that lets you create and share knowledge with your team. Unique semantic technology lets you tag such data as important dates, sales figures and customers in wiki documents or even include data from third party applications such as CRM or ERP systems. You and your teammates can reuse this data as statistics or as automatically updated lists in other documents.

SMW+ is a semantic enterprise wiki that is distributed by Ontoprise GmbH, Karlsruhe, Germany. It is well suited for organisations or teams that deal with complex and informal workflows. It offers a rich feature set to create, share, publish and re-use knowledge contained in wiki contents.

Ontoprise develops extensions to MediaWiki and carefully selects other useful ones. These extensions are integrated into SMW+ only after they pass our quality assurance procedure. Apart from team work ontology, SMW+ also includes Semantic MediaWiki, Halo Extension, Semantic Forms, WYSIWYG editor and other useful extensions.

BROWSE HELP TOPICS

- Administration and Installation of SMW+
  This section explains how to download, install and administrate SMW+

- Quickstart
  This article has a summary of all the most important user manual articles on a particular release.

- Basic Features
  This section gives details on how to work with articles, attachments and the WYSIWYG editor.

- Semantic Features
  This is a user manual for Advanced Annotation Mode, Semantic Toolbar, Ontology/Browser and many more.

Software product information

Latest release: Help:SMW+ 1.4.6

All releases: SMW+ 1.4.5, SMW+ 1.4.5 (see all releases...)

Distributed by: Ontoprise GmbH

License: GNU General Public License

Download: open here...

Now release

Fill in the name of the new release of this product and click on "add new release".

Help: add new release.
Outline

- Project Halo Overview
  - Textbooks You Can Talk To, An Analytic Encyclopedia, and Tracatable KRs for Defaults

- Deeper Dive: Crowdsourcing Data
  - Extending Wikis: Vulcan’s Semantic MediaWiki+
  - A Gallery of Semantic MediaWiki+ Applications

- Ultrapedia: An Analytic Encyclopedia
  - Ultrapedia Under the Hood
  - Ultrapedia Demo (Screenshots in Appendix)
  - The Future
What Would an Analytic Encyclopedia Look Like?

- Back to the MediaWiki roots of SMW

- Idea: Wikipedia articles merged with the power of a database
  - For Authors: tools to create more compelling articles
    - Great visualizations: charts, tables, timelines, photos, analytics
    - Always up-to-date across the Encyclopedia
    - Encourage data consistency and find data errors
    - Link in other web data sources
  - For Readers:
    - Enhanced articles and data interaction
    - Faceted navigation
    - Sophisticated queries (both standing and ad-hoc)

- Typical Problem: How do you keep the database up to date?

- Answer: Leverage the live stream of updates from millions of Wikipedia authors
  - Data is embedded in the article text, with simple ways for article authors to maintain and extend it. No DBAs.
  - Authors and readers always in the loop for merging, updating, validating, mapping
  - Data provenance is extended back to Wikipedia
Ultrapedia

Goal: Prototype a small semantic encyclopedia
- Create an semantic version of a fragment of Wikipedia
- Software is SMW with the Halo Extensions (SMW+)
- Wikipedia-based checking and corrections
- Link back to other parts of Project Halo

Ultrapedia Prototype Details
- Test domain is German cars
- ~2500 Wikipedia pages, ~40000 triples
- Private versions of Wikipedia, SMW, OB, and DBpedia hosted at wiking.vulcan.com
- Features
  - Corrections flow from Wikipedia to Ultrapedia in real time
  - Full data source tracking from Wikipedia
  - Wikipedia table ingestion and parsing
  - Feedback (user rating) loop for data
  - New visualizations for tables, charts, photos
  - External data integrated into articles
  - SPARQL-based queries
  - Derived assertions (via OntoBroker)

<table>
<thead>
<tr>
<th>Class</th>
<th>Articles</th>
<th>Infobox Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company</td>
<td>134</td>
<td>53</td>
</tr>
<tr>
<td>Person</td>
<td>93</td>
<td>57</td>
</tr>
<tr>
<td>Automobile</td>
<td>370</td>
<td>345</td>
</tr>
<tr>
<td>Auto Generation</td>
<td>1480</td>
<td>1380</td>
</tr>
<tr>
<td>Engine</td>
<td>135</td>
<td>12</td>
</tr>
<tr>
<td>Other</td>
<td>283</td>
<td>3</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>2495</strong></td>
<td><strong>1850</strong></td>
</tr>
</tbody>
</table>
Extracting Structured Data from Wikipedia

Important
Structured Data is repeated in Infobuses
Extracting Data from Wikipedia Tables

Table Data

### Engines

<table>
<thead>
<tr>
<th>Model</th>
<th>Engine</th>
<th>Power (hp, torque)</th>
<th>rpm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cayenne</td>
<td>5.598 cc (5.59 L, 3.4 L)</td>
<td>290 PS (210 kW)</td>
<td>6200</td>
</tr>
<tr>
<td>Cayenne S</td>
<td>4.806 cc (4.80 L, 293 cu in)</td>
<td>385 PS (283 kW)</td>
<td>6200</td>
</tr>
<tr>
<td>Cayenne S Transsyberia</td>
<td>4.806 cc (4.80 L, 293 cu in)</td>
<td>415 PS (309 kW)</td>
<td>6500</td>
</tr>
<tr>
<td>Cayenne GTS</td>
<td>4.806 cc (4.80 L, 293 cu in)</td>
<td>415 PS (309 kW)</td>
<td>6500</td>
</tr>
<tr>
<td>Cayenne GTS Porsche Design Edition</td>
<td>4.806 cc (4.80 L, 293 cu in)</td>
<td>415 PS (309 kW)</td>
<td>6500</td>
</tr>
<tr>
<td>Cayenne Turbo</td>
<td>4.806 cc (4.80 L, 293 cu in)</td>
<td>500 PS (370 kW)</td>
<td>6000</td>
</tr>
<tr>
<td>Cayenne Turbo B</td>
<td>4.806 cc (4.80 L, 293 cu in)</td>
<td>500 PS (370 kW)</td>
<td>6000</td>
</tr>
<tr>
<td>Cayenne Diesel</td>
<td>2.967 cc (2.967 L, 111 cu in)</td>
<td>240 PS (180 kW)</td>
<td>2400</td>
</tr>
<tr>
<td>Cayenne S Hybrid</td>
<td>3.0L supercharged V6, 3-phase synchronous electric motor</td>
<td>Petrol: 333 PS (245 kW)</td>
<td>328 hp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Electric: 52 PS (38 kW)</td>
<td>51 hp</td>
</tr>
</tbody>
</table>

### Second generation

The 2nd-generation Porsche Cayenne is expected to go on sale around April-May 2010 as a 2011 model, with an official debut at the 2010 Geneva Motor Show. In preparation for the upcoming unveiling, the Cayenne production facility in Leipzig, Germany, closed in December 2009 in order to commence the obligatory factory retooling for the new model, a process which is planned to take 2-3 months.

The first spy photos of the car were posted on the internet on the 5th of June 2008. Further spy photos, taken on June 2, 2009, and between July 2009 and January 2010 reveal a shorter, smaller Cayenne with more muscular curves, a more slanted rear window and less upright windshield, a more sloping roofline, door-mounted mirrors, smaller windows at the rear of the vehicle, headlights inspired by the Camaro GT, taillights that extend onto the car's taillight, LED daytime running lights and a vastly redesigned interior modeled after the Panamera. The new Cayenne is expected to be almost 250 kilograms lighter than the current model due to extensive use of aluminum and magnesium, making it more fuel efficient than the current Cayenne, as well as 5 centimeters shorter than the outgoing model. Due to its lower stance, the vehicle's off-road capabilities will be greatly reduced for a more performance-oriented layout and design. Diesel and hybrid variants will also be offered.

The Cayenne will again be the first of the three new SUVs from the VW group; the new Volkswagen Touareg will be 6-12 months behind, while the next-generation Audi Q7 is due in 2013. Rumored standard features of the 2011 Porsche Cayenne will include air conditioning with dual-zone climate controls, interior air filter, tilt/telescopic leather-wrapped steering wheel, wiper controls, cruise control, leather upholstery, 12-way power front seats, heated front seats, outside-temperature indicator, universal garage door opener, power liftgate, and power sunroof in the base model. The Cayenne S will add on to the climate controls, heated steering wheel, and a compass. The Cayenne GTS will add on a rearview camera, remote engine start, keyless access and start, and memory function. Finally, the most upscale Cayenne Turbo and Turbo S will add on a navigation system with voice recognition, premium sound system, 4-zone climate controls, heated rear seats, and 6-disc CD changer [11]. The new Cayenne models will also offer Porsche's new Porsche DoubleClutch (PDK) seven-speed dual clutch transmission instead of the currently-used six-speed Tiptronic S.

The Cayenne's engines are expected to receive a tuning upgrade, resulting in faster acceleration times with more horsepower and torque, as well as more powerful direct injection technology to improve efficiency. It is expected to source its V8 engines from the Panamera.

<table>
<thead>
<tr>
<th>Model</th>
<th>Engine</th>
<th>0-60 mph</th>
<th>Price (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cayenne</td>
<td>3.0L (231 kW) V6</td>
<td>7.0 seconds</td>
<td>$ 47,000</td>
</tr>
<tr>
<td>Cayenne S</td>
<td>4.5L (321 kW) V8</td>
<td>6.0 seconds</td>
<td>$ 60,000</td>
</tr>
<tr>
<td>Cayenne GTS</td>
<td>4.5L (343 kW) V8</td>
<td>5.2 seconds</td>
<td>$ 80,000</td>
</tr>
<tr>
<td>Cayenne Turbo</td>
<td>4.5L (397 kW) V8</td>
<td>4.8 seconds</td>
<td>$ 110,000</td>
</tr>
<tr>
<td>Cayenne Turbo S</td>
<td>4.5L (433 kW) V10</td>
<td>4.5 seconds</td>
<td>$ 135,000</td>
</tr>
</tbody>
</table>

See also:
- List of hybrid vehicles
Data from Wikipedia: The DBpedia Project

- The DBpedia Project ([www.dbpedia.org](http://www.dbpedia.org))
  - Publicly-funded EU Project to extract the knowledge collected by the Wikipedia community
    - Core is the Wikipedia Infobox data
    - ~700 infobox types, ~2800 property types
  - Creates a database of structured information available on the web under an open-source license

- DBpedia 3.4 dataset (Nov 09 Wikipedia)
  - ~2.9M things, ~479M triples, 80 languages
  - 282K persons, 339K places, 88K music albums, 44K films, 870K links to images, 3.8M links to relevant external web pages, 4.9M links into RDF datasets
  - 4 Classification hierarchies

- One of the largest broad knowledge bases in the world
Project Halo Enhancements to DBpedia

- **Enhanced Wikipedia parsing support**
  - Tables into semantics
  - Multiple infoboxes per page; cleaner infobox processing
  - Provenance (Wikipedia source line number) for every assertion

- **Live update stream and on-demand extraction API**

- **Data structure definition and template mapping**
  - Define an SWM ontology for template/table data
  - Data quality improvements
    - Validate input data against expected ranges, units, dimensions (“born” as a date, a city, or both?)
    - Ability to convert units (meters to kilometers to miles etc.)
  - Mappings maintained on [http://mapping.dbpedia.org](http://mapping.dbpedia.org)
    - Class and property definitions
    - Mappings from Wikipedia templates/tables to ontology classes
    - Mappings from Wikipedia template properties and table columns to ontology properties

- **Numerous bug fixes and tweaks**
Ultrapedia Prototype Data Flow

**Dynamic extraction of WP semantic data into RDF**

**Real-time feed of WP changes**
- Note most WP page changes will be text and have no semantic import

**DBpedia update stream**
- WP page text updates
- DBpedia data updates

**Ultrapedia**

**Enhanced Ultrapedia Usability**
- Familiar WP page text and layout
- Exhibit-based visualizations
- Dynamic tables/categories
- Faceted navigation
- Queries (both standing and ad-hoc)
- Wikitag-based MS Office augmentation

**Wikipedia-based Corrections**
- UP shows the user where to correct data in WP so that DBpedia will extract the correction
  - Ultrapedia exposes the data source in terms of where the data was extracted from WP
- WP changes and corrections get quickly propagated to UP

**English Wikipedia subset**

**WP updates**
- User-created page updates in Wikipedia
Demo: Ultrapedia Prototype

Ultrapedia Demo

- Domain is German cars
  - Cars, Companies, Engines, Transmissions, People, etc.
  - ~2500 pages, ~40000 triples

- An SMW-based encyclopedia
  - Similar look and feel to Wikipedia
  - Dynamic tables and charts
  - Powerful queries
  - Navigation beyond search
  - Trustworthy data source
  - Edit, discuss and rate data
  - Data is validated by the Wikipedia community

- Things to take away
  - A better Wikipedia for authors and readers
  - Interact with data as well as text
  - Monitor data with standing queries
  - External data integration via web services (EBay)

- Rapid to build
  - Software is stable
  - Most time was spent on data cleaning and new visualizations
Next Steps for Vulcan Ultrapedia

- **Needs a more natural query system**
  - There are many great SQL query formulators to adapt to ASK/SPARQL.

- **Ultrapedia for gene data is next for Project Halo, as a way for neuroscientists to collaborate on Allen Brain Atlas gene expression datasets**

- **We will also integrate with other Project Halo components (AURA/Halobook, SILK), and publish/subscribe to Linked Data**

- **What would it take to scale up to Wikipedia sizes?**
  - ~3M English articles, ~1M German articles, long tail of 264 languages
  - Key scaling factors are:
    - Table/infobox mappings: We have opened these to crowdsourcing, and have reasonable mappings for about 25% of infobox properties
    - Triplestore query times: currently ~50ms for ~40K triples, we have room to grow by a factor of 100-1000
    - Fixing data errors and table parse errors in Wikipedia
      - This is a manual process that has to be done per article
Big Concept: Crowdsourcing general RDF knowledge

- We’ve tried many ways to build large general KBs, with limited success:
  - Cycorp, Rapid Knowledge Formation, Machine Reading, and others
  - We need the eyeballs of huge numbers of people to curate and maintain the data if we are going to get to scale
- Idea: Leverage the human reference of choice, and human social curation activity

Wikipedia is already the central encyclopedia for the world; use it as a way to build a socially-curated central database for the world

Big Idea: Scale Ultrapedia to all of Wikipedia

- Ultrapedia shows many of the critical features can work (at small scale)
  - Science/technology challenges: rules, data integrity, semantic convergence, linked data, units of measure, data uncertainty, table semantics…
  - Social challenges: edit wars, neutral point of view, non-logical semantics…
Thank You