Outline

- **Background: Project Halo and Wiki Technology**
  - Problem: Building Large Science Knowledge Bases
  - Crowdsourcing Data: Semantic MediaWiki
  - An Early Attempt: Semantic Sci-Fi Movie Wiki

- **Ultrapedia**
  - A Vision of a Semantic Encyclopedia
  - Ultrapedia Under the Hood
  - Ultrapedia Demo
  - The Future
Project Halo and Wikis

- Project Halo is a staged R&D effort to build very large, computer-based 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

- Wiki is a popular way to crowd-source documents
  - Highly reliable, Internet-scale, and incredibly cheap

Can we use wiki techniques to build large knowledge bases?
A useful wiki - Wikipedia

- Wikis are tools for textual *Publication* and *Consensus*

- MediaWiki is the software for Wikipedia, Wikimedia, Wikibooks, etc.
  - Most successful Wiki package
    - High performance: 10K pages/sec served, scalability demonstrated
    - LAMP web server architecture, GPL license
    - 3 major data centers, 6-8 fulltime technical employees
  - Publication: simple distributed text authoring model
    - Wikipedia: >3.2M English articles, >270M edits, >2.5M images, #8 Alexa traffic rank in January
    - ~8.5M user accounts; ~157K active in the last month
    - 264 languages (English and German are top two)
  - Consensus achieved by global editing and rollback
    - ~17 edits/page on average (with high variance)
    - Fixpoint hypothesis, although consensus is not static
    - Gardener/admin role for contentious cases

- Wikipedia error rates are comparable to Encyclopedia Brittanica
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
How can Wikipedia be enhanced?

**Sci-Fi Movie Wiki Demo**

- **A Semantic Wiki for Science Fiction Movies from 2000 onward**
  - Look and feel are “classic” MediaWiki
  - Pages came from Wikipedia
  - Extracted semantic data from the infoboxes (via patterns, automatic)
  - Semantic UI enhancements
    - Factbox
    - Pivoting / Browsing
    - Web Services Integration
    - Sidebar Tree View

- **Things to take away:**
  - Data is now part of the wiki consensus process
  - Data can be exploited in articles
    - For mashups to other web data (Amazon in this example)
  - Data enables enhanced usability relative to Wikipedia
Outline

- **Background: Project Halo and Wiki Technology**
  - Problem: Building Large Science Knowledge Bases
  - Crowdsourcing Data: Semantic MediaWiki
  - An Early Attempt: Semantic Sci-Fi Movie Wiki

- **Ultrapedia**
  - A Vision of a Semantic Encyclopedia
  - Ultrapedia Under the Hood
  - Ultrapedia Demo
  - The Future
What Would a Powerful Encyclopedia Look Like?

- **Idea:** Wikipedia articles merged with the power of a database – enabling analysis of data

- **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)
    - Integrating and mining data inside and outside of Ultrapedia
Problem: How do you keep the data up to date?

- 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 is reviewed just like text is reviewed

Can’t do it on Wikipedia directly...
Ultrapedia: Wikipedia + a little semantics

Goal: Prototype a small semantic encyclopedia
- Create a semantic version of a part of Wikipedia
- Software is SMW and some Halo extensions
- Wikipedia-based checking and corrections

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
### Extracting Data from Wikipedia Tables

#### Table Data

**Engines**

<table>
<thead>
<tr>
<th>Model</th>
<th>Engine</th>
<th>Power (hp, torque)@rpm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cayenne</td>
<td>5.598 cc (5.598 L, 341.6 cu in) V6</td>
<td>230 PS (210 kW, 290 hp) @ 6200, 385 Nm (284 lb-ft) @ 3000</td>
</tr>
<tr>
<td>Cayenne S</td>
<td>4.806 cc (4.806 L, 293.3 cu in) V8</td>
<td>385 PS (283 kW, 500 Nm (370 lb-ft) @ 3500</td>
</tr>
<tr>
<td>Cayenne S Transsyberia</td>
<td>4.806 cc (4.806 L, 293.3 cu in) V8</td>
<td>405 PS (298 kW, 399 Nm (370 lb-ft) @ 3500</td>
</tr>
<tr>
<td>Cayenne GTS</td>
<td>4.806 cc (4.806 L, 293.3 cu in) V8</td>
<td>405 PS (298 kW, 399 Nm (370 lb-ft) @ 3500</td>
</tr>
<tr>
<td>Cayenne GTS Porsche Design Edition</td>
<td>4.806 cc (4.806 L, 293.3 cu in) V8</td>
<td>405 PS (298 kW, 399 Nm (370 lb-ft) @ 3500</td>
</tr>
<tr>
<td>Cayenne Turbo</td>
<td>4.806 cc (4.806 L, 293.3 cu in) turbo V8</td>
<td>510 PS (370 kW, 490 Nm @ 6600, 700 Nm (520 lb-ft) @ 2250-4500</td>
</tr>
<tr>
<td>Cayenne Turbo S</td>
<td>4.806 cc (4.806 L, 293.3 cu in) twin turbo V8</td>
<td>550 PS (400 kW, 540 Nm @ 6600, 750 Nm (550 lb-ft) @ 2250-4500</td>
</tr>
<tr>
<td>Cayenne Diesel</td>
<td>2.967 cc (2.967 L, 111.1 cu in) turbo V6</td>
<td>240 PS (180 kW, 240 Nm) @ 5500, 550 Nm (410 lb-ft) @ 2000</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, 328 hp), 439 Nm (324 lb-ft) @ 2900-5300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Electric: 52 PS (38 kW, 51 hp), 330 Nm (220 lb-ft)</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 retouching 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.[9] Further spy photos, taken on June 2, 2009[10] and between July 2009 and January 2010 reveal a 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 Cayenne GT, taillights that extend onto the car’s tailgate, 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 diesel and hybrid variants will also be offered.

The Cayenne will again be the first of the three new SUVs of 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, tilting/telescopic leather-wrapped steering wheel, radio 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 three 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 system. 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 Doppelkupplung (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 (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cayenne</td>
<td>3.6L (231 kW) V6</td>
<td>7.0 seconds</td>
<td>$ 57,000</td>
</tr>
<tr>
<td>Cayenne S</td>
<td>3.6L (211 kW) V6</td>
<td>6.0 seconds</td>
<td>$ 50,000</td>
</tr>
<tr>
<td>Cayenne GTS</td>
<td>3.6L (343 kW) V8</td>
<td>5.2 seconds</td>
<td>$ 80,000</td>
</tr>
<tr>
<td>Cayenne Turbo</td>
<td>3.6L (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 related 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 3.5 dataset (Mar 10 Wikipedia)**
  - ~3.4M things, ~1B triples, 92 languages
  - 312k persons, 413k places, 94k music albums, 49k films, 841k links to images, 5.1M links to relevant external web pages, 9.4M links into RDF datasets
  - 3 Classification hierarchies

- **A large and broad knowledge base in the world**
Skyscrapers in China
higher than 50 stories
built before year 2000

{{#ask:
[[Category:Skyscrapers]]
[[Located in::China]]
[[Floor count::>50]]
[[Year built::<2000]]
... (output format)
}}
**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

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

English Wikipedia subset

DBpedia update stream
• WP page text updates
• DBpedia data updates

WP updates
• User-created page updates in Wikipedia

Dynamic extraction of WP semantic data into RDF

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
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
  - With dynamic tables
  - Flexible and powerful queries
  - Many visualization methods
  - Trustworthy data source
  - Edit, discuss and rate data

- **Things to take away**
  - A better Wikipedia
  - Some pages are Ultrapedia only
  - External data integration via web services (EBay)
  - Real-time data updates

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

- Even better data extraction system
- Needs a natural, easy-to-use query system
- Wikipedia for neuroscience knowledge is next
  - Size is about 10x German cars
- We will also integrate with other Project Halo components
- What would it take to scale up to Wikipedia?
  - ~3M English articles, ~1M German articles, long tail of 264 languages
  - Key scaling factors are:
    - Table/infobox mappings: We have opened these to crowdsourcing
    - 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 is done per article
SMW Contributors

Paul Allen
Free University of Berlin
John Chu
Mark Greaves
Benjamin Grososf
Dave Gunning
Rob Kaplan
Yaron Koren
University of Karlsruhe
Dave Moore
Ontoprise GmbH
TeamMersion LLC
Jesse Wang
Thank You