Leveraging Semantic Structure to Improve Retrieval: Restructuring Wikipedia using Topic Maps and RDF Ontologies

•••

Wonhong Jang and Sam Oh
Sungkyunkwan University, iSchool & Data Science

Aim of Study

- **Restructuring Wikipedia data to achieve a better browsing experience employing RDF/OWL and Topic Maps ontologies.**

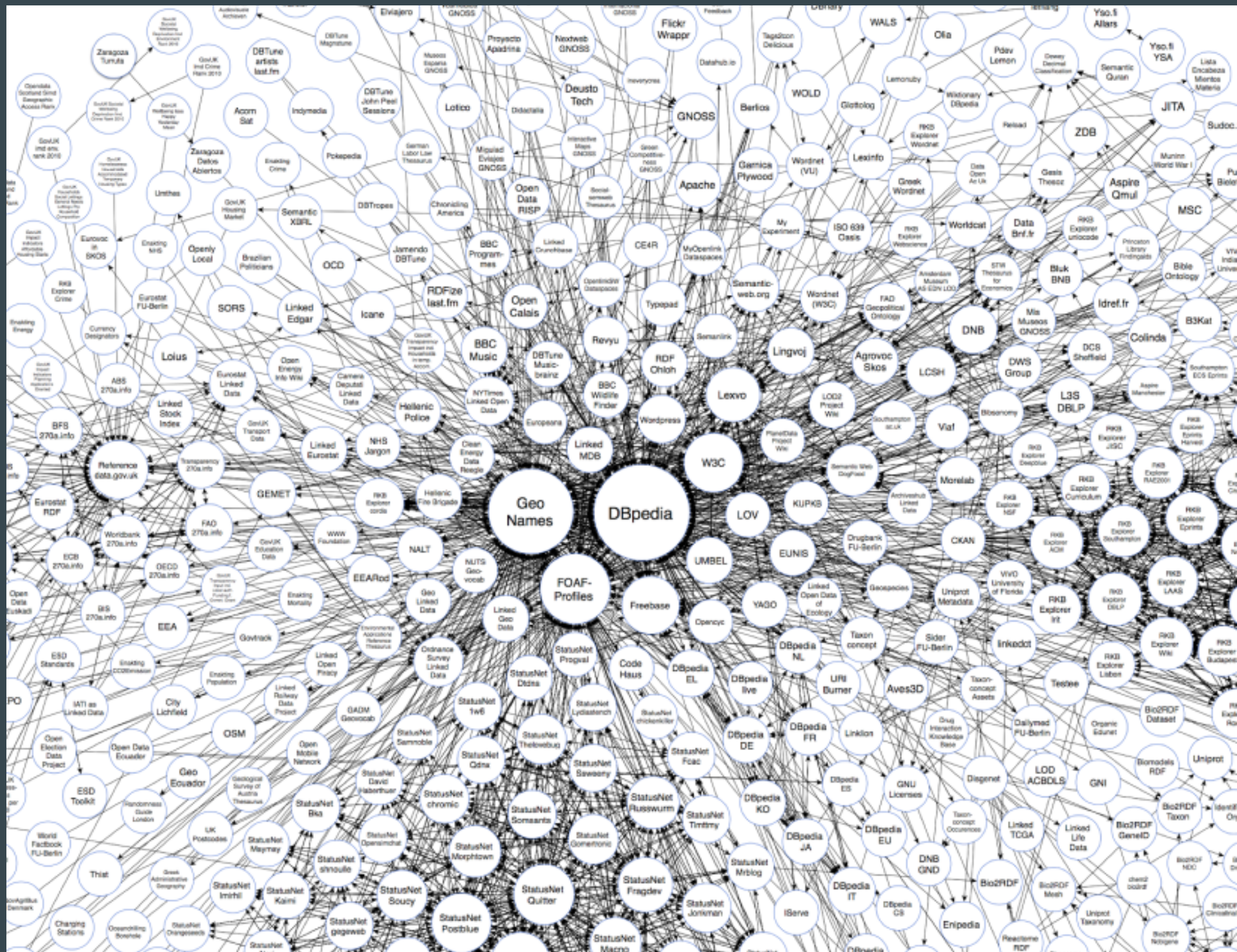
- **Conducting objective and subjective evaluations of how ontology-enabled Wikipedia differ from the existing one**

Research Variables

- **Independent Variables**
  - The Current Wikipedia System
  - RDF/OWL-based Wikipedia System
  - Topic Maps-based Wikipedia System
- **Dependent Variables**
  - The Objective Measurements
    - Retrieval time, Retrieval iterations, and Number of pages viewed
  - The Subjective Measurements
    - Degree of Accuracy, Usefulness, and Satisfaction

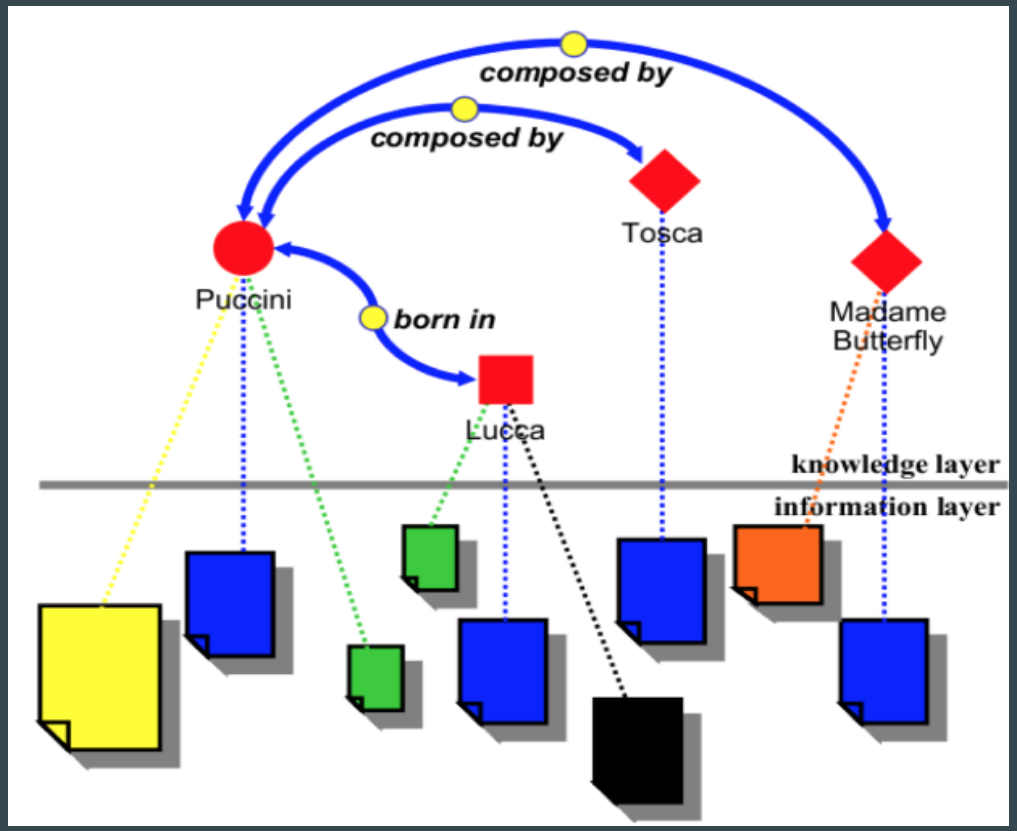
DBpedia

DBpedia is a project dedicated to extracting structured information from Wikipedia data and making it possible to access it on the Web



Topic Maps

- Topic Maps is one of the key description languages used when constructing ontologies, which in turn serve as a key component of the Semantic Web
- It was approved as an ISO standard in 2000



Topic Maps Model
(Opera Topic Map)

RDF/OWL

- Standardized by the W3C, RDF(Resource Description Framework) is one of the basic languages of the Semantic Web which constitutes a framework for describing information resources or metadata for the Web. It is expressed in the triple form in which subject, predicate, and object form a single set.

- OWL is an extension of RDF and RDF Schema created for the purpose of ontology construction, through which a wide variety of inference engines have been built.

Semantic Query Languages

- **TOLOG** is a semantic query language for Topic Maps formulated by combining the advantages of Prolog and SQL. Just as SQL can be used to obtain various query results from a RDBMS, so can TOLOG be used to glean a similarly diverse set of query results from a Topic Maps ontology and to construct complex systems through query design

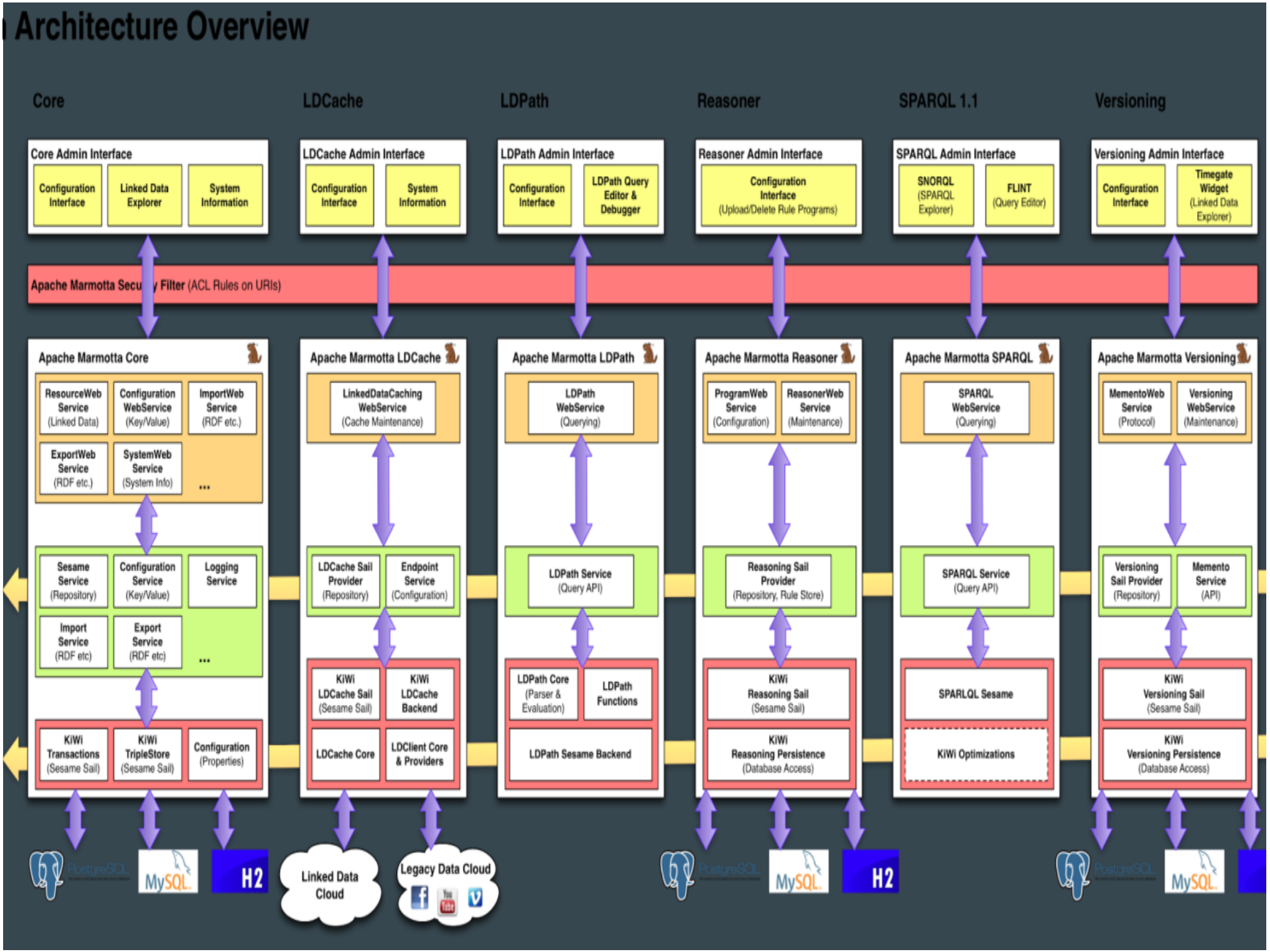
- **SPARQL** is the query language for RDF/OWL technology. It was standardized by the World Wide Web Consortium and became a W3C Recommendation in 2008. An advanced form of RDQL, a precursor query language for RDF graphs, SPARQL enables the modeling of questions and answers in the Semantic Web environment

Implementation of Test Systems

- Data
  - Ingested DBpedia (already in RDF/OWL) and aligned them to LMF data structure.
  - Ingested DBPedia into Ontopoly TM Editor, enhanced relationships using TM associations, and aligned enhanced data to OKS data structure.
- Platforms
  - LMF(Linked Media Framework) for RDF/OWL-based Wikipedia
  - OKS(Ontopia Knowledge Suite) for Topic Maps-based Wikipedia
- Web Interfaces
  - Employed TOLOG and SPARQL to improve retrieval functionality of the existing Wikipedia system

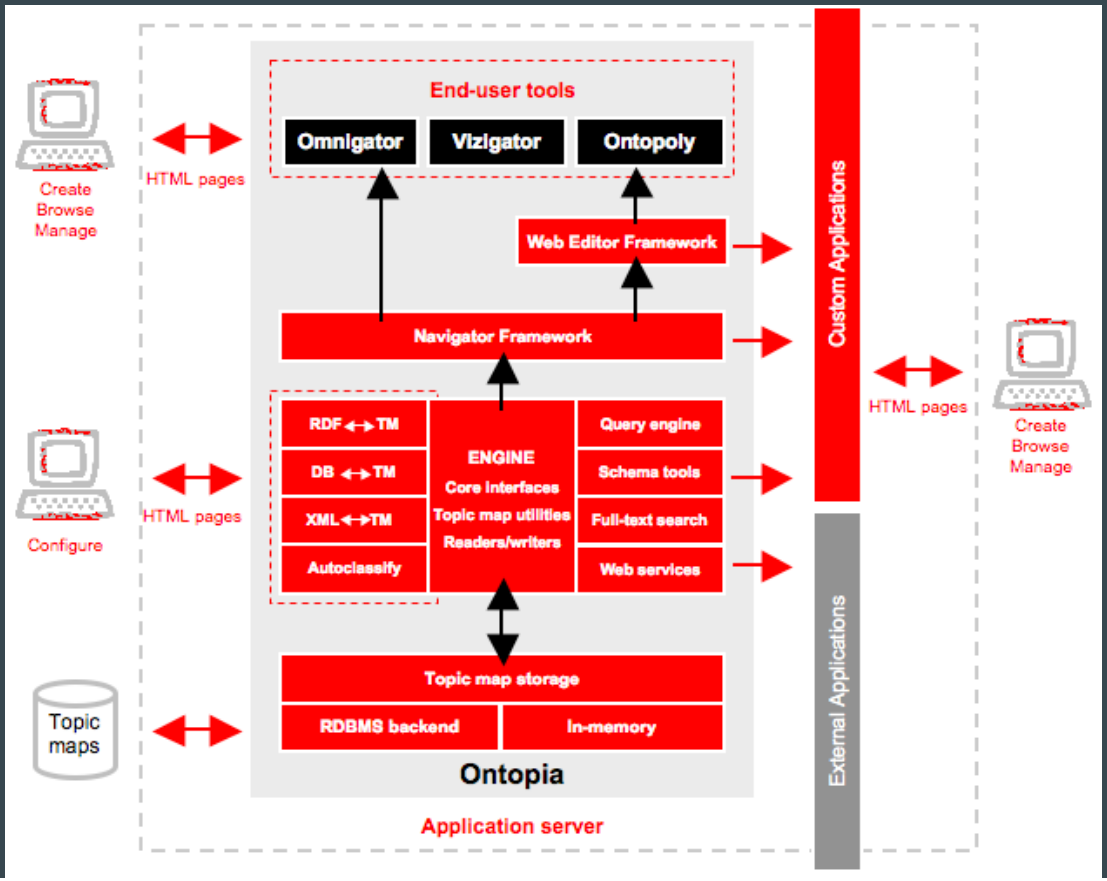
Implementation of Test Systems

- LMF(Linked Media Framework)
  - Easy-to-setup application platform that bundles together some key open source modules to build RDF/OWL based system
  - The name “LMF” is now changed into “Apache marmotta”


Architecture Overview
Core
LDCache
LDPath
Reasoner
SPARQL 1.1
Versioning
Core Admin Interface
Configuration Interface
Linked Data Explorer
System Information
LDCache Admin Interface
Configuration Interface
System Information
LDPath Admin Interface
Configuration Interface
LDPath Query Editor & Debugger
Reasoner Admin Interface
Configuration Interface (Upload/Delete Rule Programs)
SPARQL Admin Interface
SNORQL (SPARQL Explorer)
FLINT (Query Editor)
Versioning Admin Interface
Configuration Interface
Timegate Widget (Linked Data Explorer)
Apache Marmotta Security Filter (ACL Rules on URIs)
Apache Marmotta Core
ResourceWeb Service (Linked Data)
Configuration WebService (Key/Value)
ImportWeb Service (RDF etc.)
ExportWeb Service (RDF etc.)
SystemWeb Service (System Info)
...
Sesame Service (Repository)
Configuration Service (Key/Value)
Logging Service
Import Service (RDF etc)
Export Service (RDF etc)
...
KiWi Transactions (Sesame Sail)
KiWi TripleStore (Sesame Sail)
Configuration (Properties)
Apache Marmotta LDCache
LinkedDataCaching WebService (Cache Maintenance)
LDCache Sail Provider (Repository)
Endpoint Service (Configuration)
KiWi LDCache Sail (Sesame Sail)
KiWi LDCache Backend
LDCache Core
LDClient Core & Providers
Apache Marmotta LDPath
LDPath WebService (Querying)
LDPath Service (Query API)
LDPath Core (Parser & Evaluation)
LDPath Functions
LDPath Sesame Backend
Apache Marmotta Reasoner
ProgramWeb Service (Configuration)
ReasonerWeb Service (Maintenance)
Reasoning Sail Provider (Repository, Rule Store)
KiWi Reasoning Sail (Sesame Sail)
KiWi Reasoning Persistence (Database Access)
Apache Marmotta SPARQL
SPARQL WebService (Querying)
SPARQL Service (Query API)
SPARLQL Sesame
KiWi Optimizations
Apache Marmotta Versioning
MementoWeb Service (Protocol)
Versioning WebService (Maintenance)
Versioning Sail Provider (Repository)
Memento Service (API)
KiWi Versioning Sail (Sesame Sail)
KiWi Versioning Persistence (Database Access)
PostgreSQL
MySQL
H2
Linked Data Cloud
Legacy Data Cloud
PostgreSQL
MySQL
H2
PostgreSQL
MySQL

Implementation of Test Systems

- OKS(Ontopia Knowledge Suite)
  - Open source framework for building and deploying Topic Map-based system



Objective Measurements: Detail

Scope	Point of Evaluation	Description	Method of Recording	Unit of Recording
Objective Measurements	Retrieval Time	How much time was spent before arriving at the final retrieval results?	Time taken	each query
	Retrieval Iterations	How many keywords were used before arriving at the final retrieval results?	Keyword counts	each query
	Page Views	How many pages were viewed before arriving at the final retrieval results?	Pages viewed	each query

Subjective Measurements: Detail

Scope	Point of Evaluation	Survey Question	Unit of Recording
Subjective Measurements	Accuracy	I was able to obtain accurate answers to the given tasks by using this system.	Query type
		I was able to obtain accurate results to the keywords inputted into the search engine.	Query type
	Usability	This system provides a variety of different search paths.	Query type
		This system is easy to navigate.	Query type
	Usefulness	I was able to obtain a quick answer to the given tasks by using this system.	Query type
		I was able to obtain additional information while searching for answers to the given tasks.	Query type
	User Satisfaction	I am satisfied with the retrieval results obtained by using this system.	Individual systems
		I am satisfied with this system overall.	Individual systems

Query Types

Query Type	Query Template	Example Query
Simple Query	Search for (Sports Event) held in (Year).	Search for the host city of 1997 East Asian Games.
	Search for the year of birth and year of death of (Person).	Search for birth and death year of Charles Darwin.
	Search for information on the members of (Person)'s family.	Search for information on members of Charles Darwin's family.
Composite Query	Search for the title and publication year of the book written by (Author) in (Year).	Search for the title and publication year of the book written by George Orwell in 1948.
	Search for the company which published (Book Title) in (Year). Then search for a list of the company's representative publications.	Search for the company which published Bill Clinton's autobiography *My Life* in 2004. Then search for a list of the company's representative publications.
	Search for the address of the place where (Person) lived until his/her death.	Search for the house address where Albert Einstein lived until his death.
User Query	Each subject conducts his/her own queries.	

Experimental Procedure

Stage	Description
Step 1	**Experiment explained:** All test subjects were provided with an explanation of the purpose and methodology of the study. They were made aware of the fact that each step of the experiment would be recorded or filmed and were asked to grant permission for this process. All experiments were carried out in the same physical space using the same systems.
Step 2	**Subject assignment(3 groups):** The 21 test subjects were equally divided into **three groups** and asked to begin the experiment with a different system, thereby eliminating the negative effects of familiarizing oneself with the test systems in a particular order. Order effects are cancelled out by switching orders.
Step 3	**Starting a search:** All test subjects were informed of the particulars of the experiment and their permission obtained before they were introduced to the new system and provided with two sample search queries.
Step 4	**Objective measures recorded:** The amount of time spent per search query was recorded. Video of each search was taken from the moment when the test subject first inputted keywords in response to the search queries to the time when the test subject was presented with his or her search results.
Step 5	**Subjective measures recorded:** The test subjects were asked to fill out surveys on all three systems following the completion of the search queries for each system.

Methods of Data Collection

Methods	Description	Information Collected
Screen Recording	Used a screen capture program to record footage of the test subjects' system use in real time for later analysis	-Retrieval time - Retrieval Iterations -Page views
Survey	Referred to the DeLone and McLean Model of Information System Success regarding its standards of information quality, information use, and degree of user satisfaction before providing surveys to test subjects	-System accuracy -System usefulness -User satisfaction

Results: Objective Measures - **Simple Query**

- One-Way ANOVA and Post-hoc Results on **Simple Query** Retrieval Time, Number of Retrievals, and Page Views

Hypothesis	F Value	Pr>F	Hypothesis	Post-hoc Test			
				Scheffe Grouping	Mean	N	System
A-1. The amount of time needed for the retrieval will vary	7.48	0.0013	Accepted	A	199.71	21	WIKI
				B	156.62	21	RDF/OWL
				B	144.9	21	Topic Map
A-2. The number of retrieval iterations will vary	0.86	0.428	Rejected	Post-hoc test not needed			
A-3. The number of pages viewed will vary	4.51	0.0149	Accepted	A	5.6667	21	WIKI
				B	4	21	RDF/OWL
				B	4.0952	21	Topic Map

Results: Objective Measures - **Composite Query**

- One-Way ANOVA and Post-hoc Results on **Composite Query** Retrieval Time, Number of Retrievals, and Page Views

Hypothesis	F Value	Pr>F	Hypothesis	Post-hoc Test			
				Scheffe Grouping	Mean	N	System
B-1. The amount of time needed for the retrieval will vary	20.3	<.0001	Accepted	A	419.05	21	WIKI
				B	245.71	21	RDF/OWL
				B	234.1	21	Topic Map
B-2. The number of retrieval iterations for composite queries will vary	4.41	0.0163	Accepted	A	6.9048	21	WIKI
				A	6.381	21	RDF/OWL
				B	5.2857	21	Topic Map
B-3. The number of pages viewed for composite queries will vary	1.78	0.1778	Rejected	Post-hoc test not needed			

Results: Objective Measures - **Subject Queries**

- One-Way ANOVA and Post-hoc Results on Retrieval Time, Number of Retrievals, and Number of Page Views for **Test Subject Queries**

Hypothesis	F Value	Pr>F	Hypothesis	Post-hoc Test			
				Scheffe Grouping	Mean	N	System
C-1. The amount of time needed for the retrieval will vary	0.33	0.7196	Rejected	Post-hoc test not needed			
C-2. The number of retrieval iterations will vary	0.11	0.8974	Rejected	Post-hoc test not needed			
C-3. The number of pages viewed will vary	0.33	0.7222	Rejected	Post-hoc test not needed			

Results: Subjective Measures - **Simple Query**

- One-Way ANOVA and Post-hoc Results on the Accuracy, Usability, and Usefulness of **Simple Query** Results

Hypothesis	F Value	Pr>F	Hypothesis	Post-hoc Test			
				Scheffe Grouping	Mean	N	System
D-1. The degree of accuracy felt by subjects will vary	6.54	0.002	Accepted	A	8.5238	42	Topic Map
				A	8.119	42	RDF/OWL
				C	7.7857	42	WIKI
D-2. The degree of usability felt by subjects will vary	40.59	<.0001	Accepted	A	8.5	42	Topic Map
				B	7.7381	42	RDF/OWL
				C	6.6667	42	WIKI
D-3. The degree of usefulness felt by the test subjects will vary	15.48	<.0001	Accepted	A	8.6905	42	Topic Map
				B	7.9048	42	RDF/OWL
				B	7.4524	42	WIKI

Results: Subjective Measures - **Composite Query**

- One-Way ANOVA and Post-hoc Results on the Accuracy, Usability, and Usefulness of **Composite Query** Results

Hypothesis	F Value	Pr>F	Hypothesis	Post-hoc Test			
				Scheffe Grouping	Mean	N	System
E-1. The degree of accuracy felt by the subjects will vary	25.43	<.0001	Accepted	A	8.2857	42	Topic Map
				B	7.4524	42	RDF/OWL
				C	6.7143	42	WIKI
E-2. The degree of usability felt by the test subjects will vary	44.05	<.0001	Accepted	A	8.3095	42	Topic Map
				B	7.3333	42	RDF/OWL
				C	6.2381	42	WIKI
E-3. The degree of usefulness felt by the test subjects will vary	21.04	<.0001	Accepted	A	8.4048	42	Topic Map
				B	7.5	42	RDF/OWL
				C	6.7619	42	WIKI

Results: Subjective Measures - **Subject Queries**

- One-Way ANOVA and Post-hoc Results on the Accuracy, Usability, and Usefulness of the Results of the **Test Subject Queries**

Hypothesis	F Value	Pr>F	Hypothesis	Post-hoc Test			
				Scheffe Grouping	Mean	N	System
F-1. The degree of accuracy felt by the test subjects will vary	10.24	<.0001	Accepted	A	8.3333	42	Topic Map
				A	7.8571	42	RDF/OWL
				B	7.381	42	WIKI
F-2. The degree of usability felt by the test subjects will vary	33.44	<.0001	Accepted	A	8.4524	42	Topic Map
				B	7.4286	42	RDF/OWL
				C	6.6667	42	WIKI
F-3. The degree of usefulness felt by the test subjects will vary	19.02	<.0001	Accepted	A	8.5238	42	Topic Map
				B	7.881	42	RDF/OWL
				C	7.0476	42	WIKI

Conclusion

- **Objective Measures**
  - Simple queries: Retrieval time, Pages viewed
  - Composite queries: Retrieval time, Retrieval iterations
  - Subject queries: None

- **Subjective Measures**
  - Simple queries: Accuracy, Usability, and Satisfaction (all variables)
  - Composite queries: Accuracy, Usability, and Satisfaction (all variables)
  - Subject queries: Accuracy, Usability, and Satisfaction (all variables)

Thank you for your attention!!
Any questions?