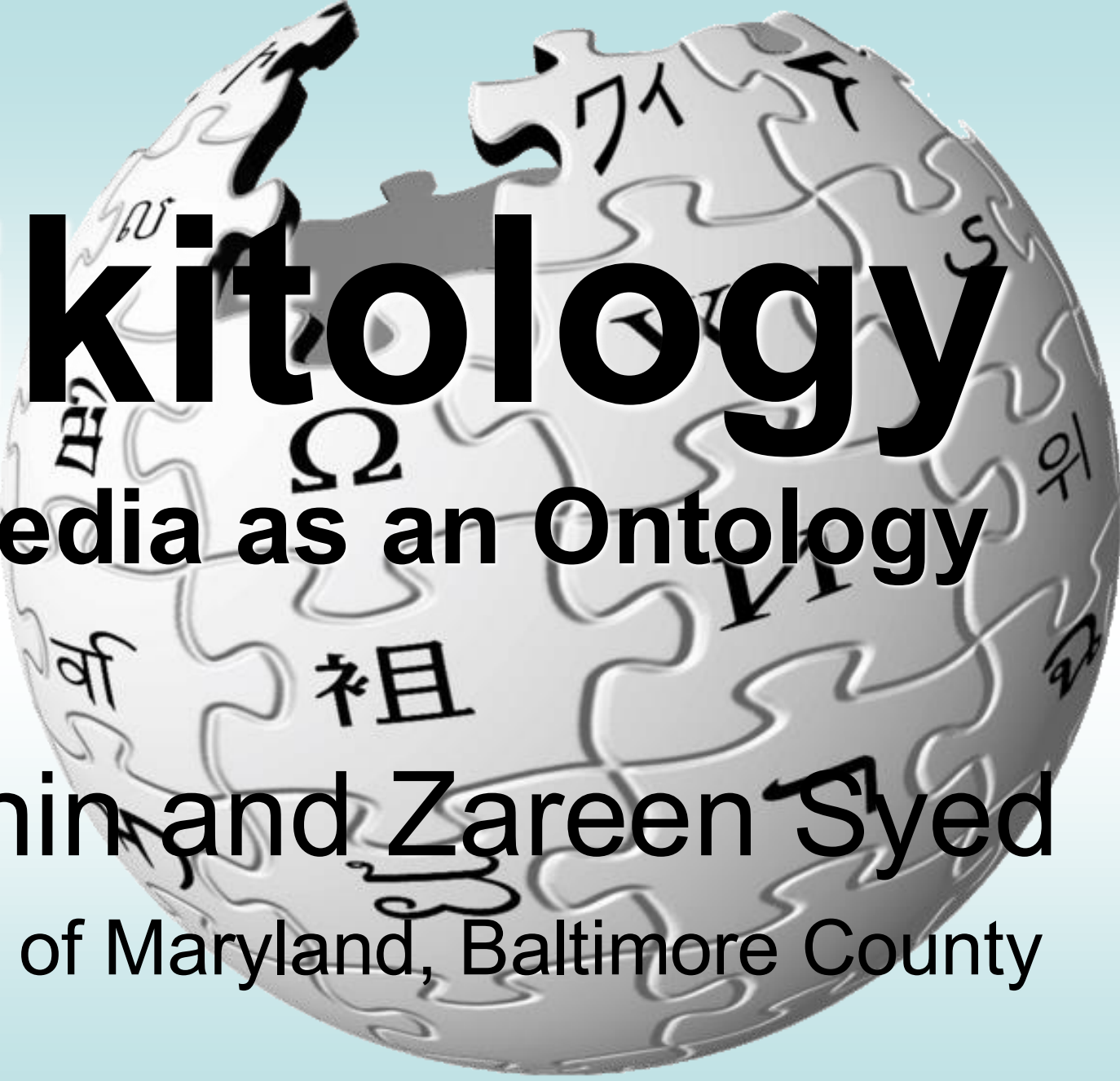


# Wikitalogy

A globe composed of interlocking puzzle pieces. Each piece features a different character or symbol from various alphabets, including Latin, Greek, and Chinese. The globe is set against a light blue background.

## Wikipedia as an Ontology

Tim Finin and Zareen Syed

University of Maryland, Baltimore County

# Outline

- Introduction and motivation
- Wikipedia 101
- Experiments
- Evaluation
- Next steps
- Conclusion

# Overview

- **Problem:** describe what an analyst has been working on to support collaboration
- **Idea:** track documents she reads and map these to terms in an ontology, aggregate to produce a short list of topics
- **Approach:** use Wikipedia articles as ontology terms, use document-article similarity for the mapping, and spreading activation for aggregation

# What's a document about?

- Two common approaches:
  - (1) Select words and phrases using TF-IDF that characterize the document
  - (2) Map document to a list of terms from a controlled vocabulary or ontology
- (1) is flexible and does not require creating and maintaining an ontology
- (2) can tie documents to a rich knowledge base

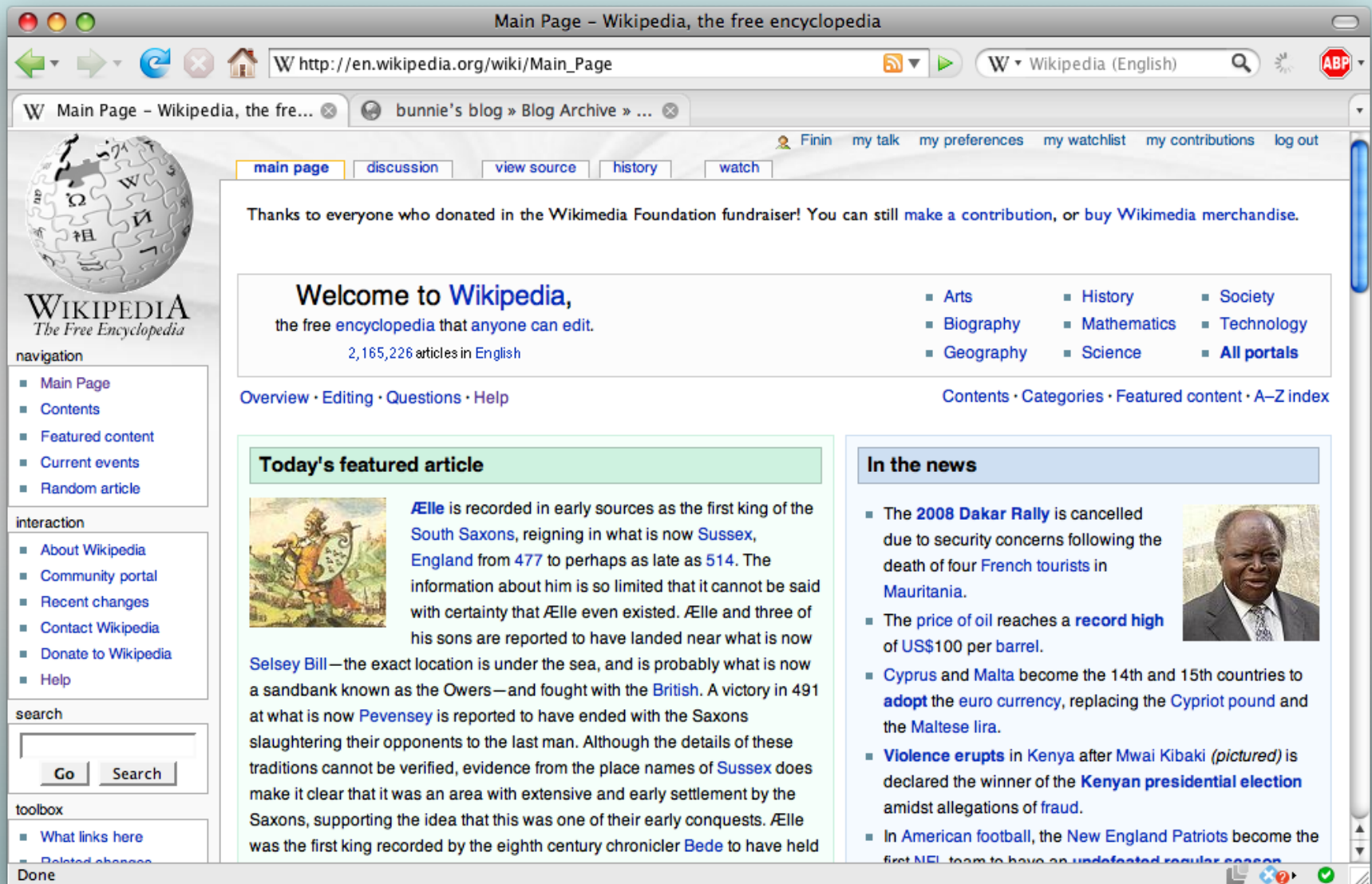
# Wikitology !

- Using Wikipedia as an ontology offers the best of both approaches
  - Each article is a concept in the ontology
  - Terms linked via Wikipedia's category system and inter-article links
- It's a consensus ontology created, kept current and maintained by a diverse community
- Overall content quality is high

# Wikitology features

- Terms have unique IDs (URLs) and are “self describing” for people
- Several underlying graphs provide structure: categories, article links
- Article history contains useful meta-data (e.g., for trust)
- External sources provide more info (e.g., Google’s pagerank)
- Some of the data available in structured form, e.g., in RDF from DBpedia

# Wikipedia 101



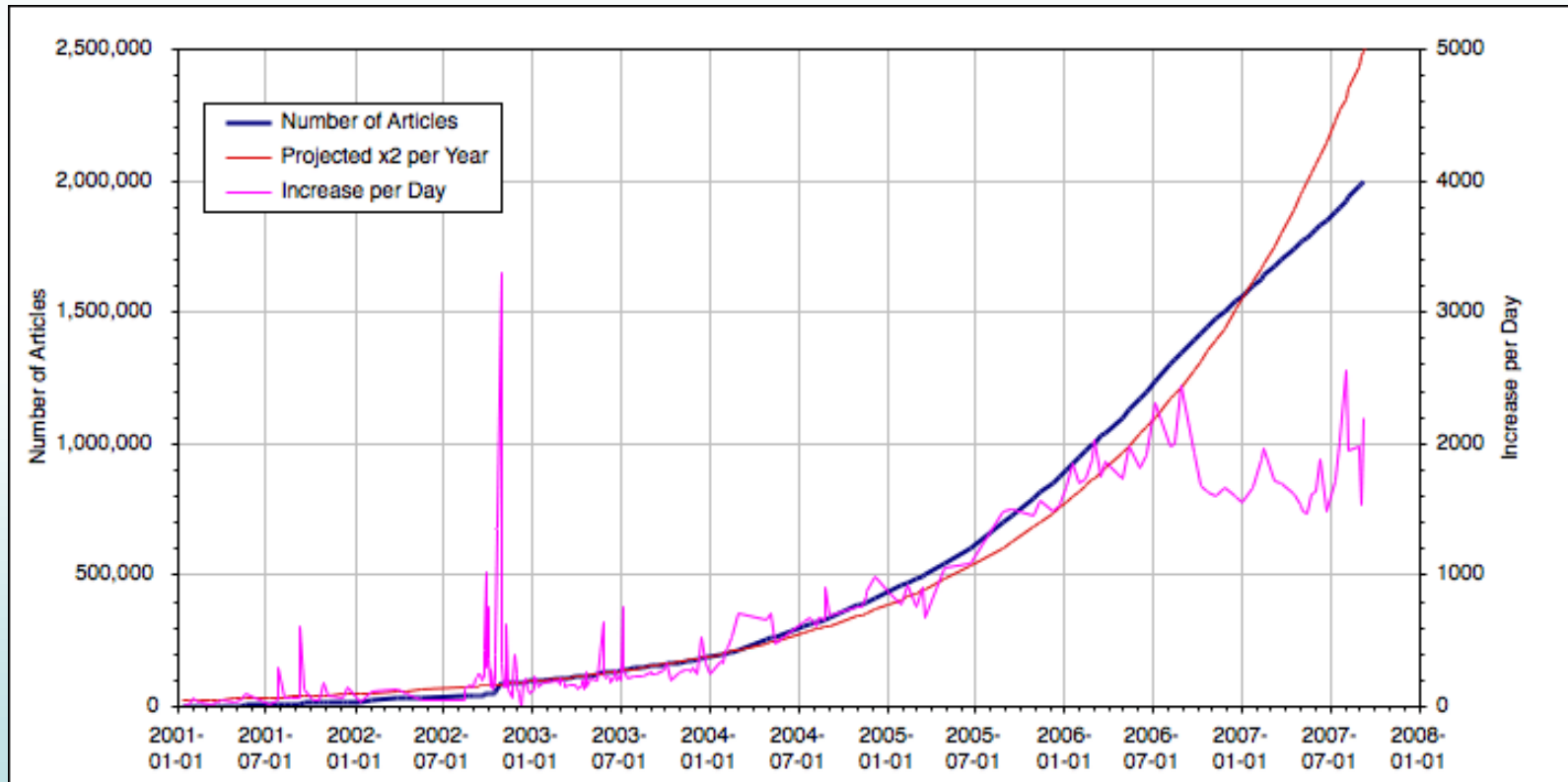
# Wikipedia history

- Started January 2001 to complement the peer-reviewed [Nupedia](#) project
- Based on Ward Cunningham's Wiki idea (wiki wiki is Hawaiian for *quick!*)



# Wikipedia's size and growth

- 9.25M articles in 253 languages, 1.4B words
- English: 2.2M articles, 940M words -- largest encyclopedia ever assembled
- 6.2M registered users, 192M edits



# Wikipedia data in RDF



The screenshot shows a web browser window with the address bar displaying "http://dbpedia.org/". The page title is "wiki.dbpedia.org : About". The DBpedia logo is prominently displayed at the top left. A sidebar on the left contains a list of navigation links: "About / News", "Datasets", "Online Access", "Downloads", "ChangeLog", "Interlinking", "Use Cases", "Framework", "Publications", "Community", "Blog", and "Credits". The main content area features a paragraph describing DBpedia as a community effort to extract structured information from Wikipedia. Below this, a red header labeled "News" introduces a list of recent updates, including a presentation at ISWC, Cyc linkage information, and flickr photo collection links.

wiki.dbpedia.org : About

W. http://dbpedia.org/

dbpedia



**About / News**

- Datasets
- Online Access
- Downloads
- ChangeLog
- Interlinking
- Use Cases
- Framework
- Publications
- Community
- Blog
- Credits

**DBpedia** is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia and to link other datasets on the Web to Wikipedia data.

**News**

[DBpedia-Presentation at ISWC](#)  
Sören presented today the paper "DBpedia: A Nucleus for a Web of Open Data" at International Semantic Web Conference in Busan, Korea. You can view the slides at <http://www.informatik.uni-leipzig.de/~auer/presentations/dbpedia/>

[DBpedia-Cyc linkage](#)  
The commonsense knowledge base Cyc or OpenCyc (when compared to DBpedia) seems to follow a rather top-down approach ? first more abstract concepts and entities are represented and later Cyc started to include also more domain knowledge. This seems to be reasonable, since domain knowledge changes faster and there is much more of it. On [...]

[flickr photo collection links added to DBpedia](#)  
Christian Becker (Freie Universität Berlin) has implemented a wrapper around flickr which generates photo collections depicting DBpedia concepts. See flickr wrappr for details. We have interlinked all DBpedia concepts with the corresponding photo collections. You can now use any Semantic Web browser to navigate from a DBpedia concept to flickr photos depicting it by following the `dbpedia:hasPictureCollection` property. [...]

Done

# Populating Freebase KB

Alan Turing : Freebase

http://www.freebase.com/view/alan\_turing

wiki.dbpedia.org : About Alan Turing : Freebase

freebase<sup>TM</sup>  
alpha

Keyword search Freebase Search

Home Data Apps Discuss Help Welcome back, finin. Not you? Sign out.

## Alan Turing

Discuss "Alan Turing" Show Empty Fields



Image 1 of 1

- Types:** Person (People), Book Subject (Publishing), Computer Scientist (Computers), Deceased Person (People), Influence Node (mikelove's types), Computer Designer (Computers)
- Date of death:** Jun 7, 1954
- Place of death:** Manchester, England
- Cause of death:** Cyanide Poisoning, Suicide
- Gender:** Male
- Date of Birth:** Jun 23, 1912
- Place of Birth:** London
- Profession:** Computer Scientist, Mathematician, Scientist, Philosopher
- Religion:** Atheism
- Parents:** Ethel Stoney Turing, Julius Mathison Turing

### Description

Alan Mathison Turing, OBE, FRS (23 June 1912 – 7 June 1954) was an English mathematician, logician, and cryptographer. Turing is often considered to be the father of modern computer science. Turing provided an influential formalisation of the concept of the algorithm and

### Page History

Created by Metaweb Oct 22, 2006  
Last edited by mikelove Nov 1, 2007

### Books About This Topic

Cryptonomicon

### Education

Princeton University • 1937 • 1938 • Ph.D.  
King's College, Cambridge • 1931 • 1934

### Quotations

Mathematical reasoning may be regarded...

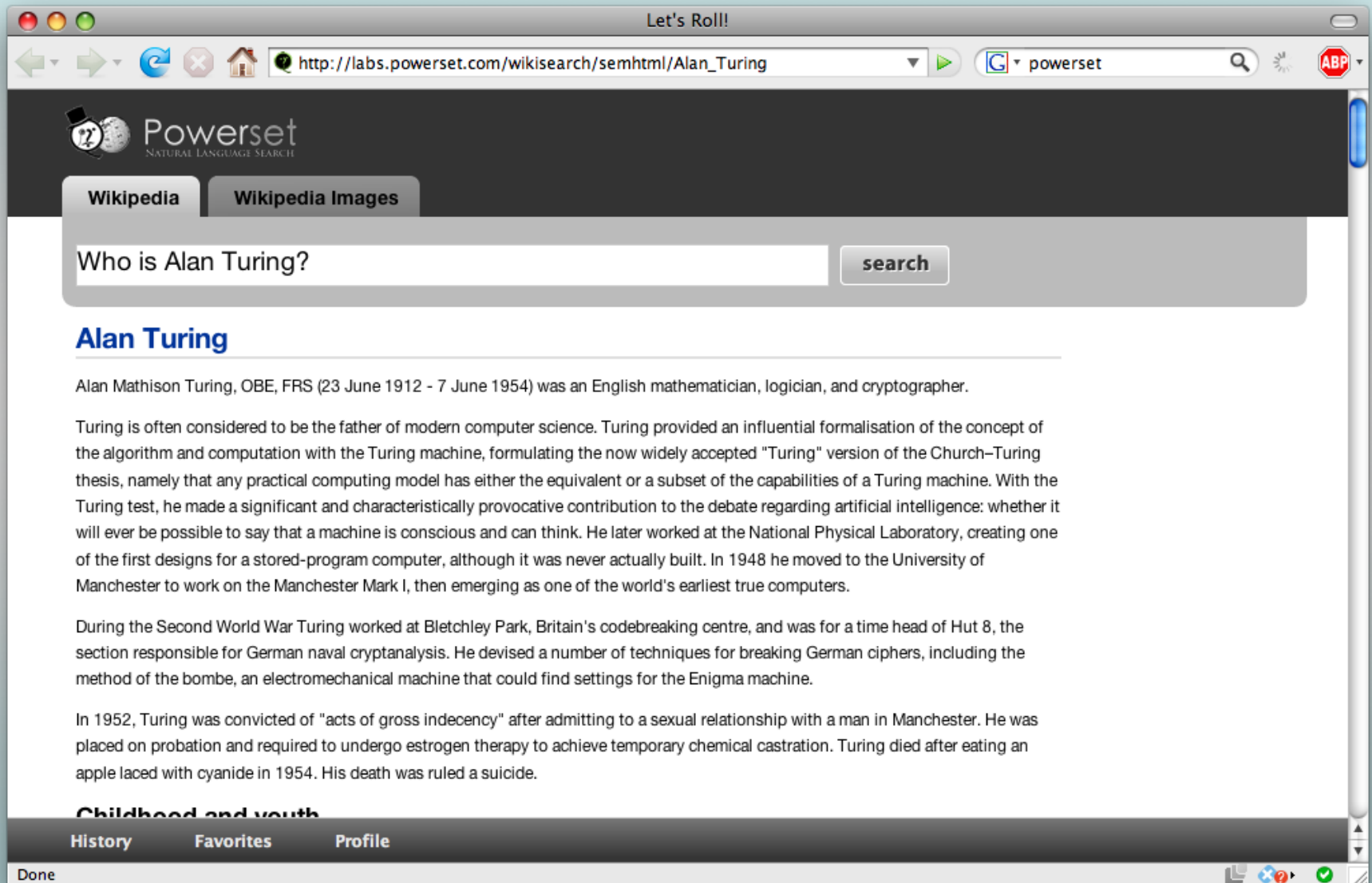
No, I'm not interested in developing a powerful brain...

Science is a differential equation.  
Religion is a boundary condition.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

Done

# Populating Powerset's KB



# AskWiki uses Wikipedia for QA

Main Page - AskWiki

http://askwiki.com/AskWiki/index.php/Main\_Page

aks wiki

76.114.193.200 talk for this ip log in / create account

article discussion view source history

## Main Page

Welcome to **AskWiki Beta**  
The Wikipedia Answer-Based Search Engine

**Enter Your Question Below:**

Who is alan turing?

**Answer to Your Question: "who is alan turing?"**

Alan Mathison Turing, OBE, FRS (23 June 1912 - 7 June 1954) was an English mathematician, logician, and cryptographer.

ASK WIKI

navigation

- [Main Page](#)
- [Community portal](#)
- [About AskWiki](#)
- [How It Works](#)
- [Top Categories](#)
- [Random Q&A](#)
- [Correct Answers](#)
- [Incorrect Answers](#)
- [Uncertain Answers](#)

askwiki engine

article search

Done

# With sometimes surprising results

The screenshot shows a web browser window titled "Main Page - AskWiki". The address bar displays "http://askwiki.com/AskWiki/index.php/Main\_Page". The page features a navigation menu on the left with links like "Main Page", "Community portal", and "About AskWiki". The main content area has tabs for "article", "discussion", "view source", and "history". A large banner reads "Welcome to AskWiki Beta" and "The Wikipedia Answer-Based Search Engine". Below this, a search bar contains the text "Does God exist?" and a button labeled "AskWiki". The search results show the question "Answer to Your Question: 'Does God exist?'" followed by a paragraph: "Some definitions of God are so nonspecific that it is certain that something exists that meets the definition; while other definitions are apparently self-contradictory." The browser's status bar at the bottom shows "Done".

Main Page - AskWiki

http://askwiki.com/AskWiki/index.php/Main\_Page

76.114.193.200 talk for this ip log in / create account

article discussion view source history

## Main Page

Welcome to **AskWiki Beta**  
The Wikipedia Answer-Based Search Engine

**Enter Your Question Below:**

Does God exist? AskWiki

**Answer to Your Question: "Does God exist?"**

Some definitions of God are so nonspecific that it is certain that something exists that meets the definition; while other definitions are apparently self-contradictory.

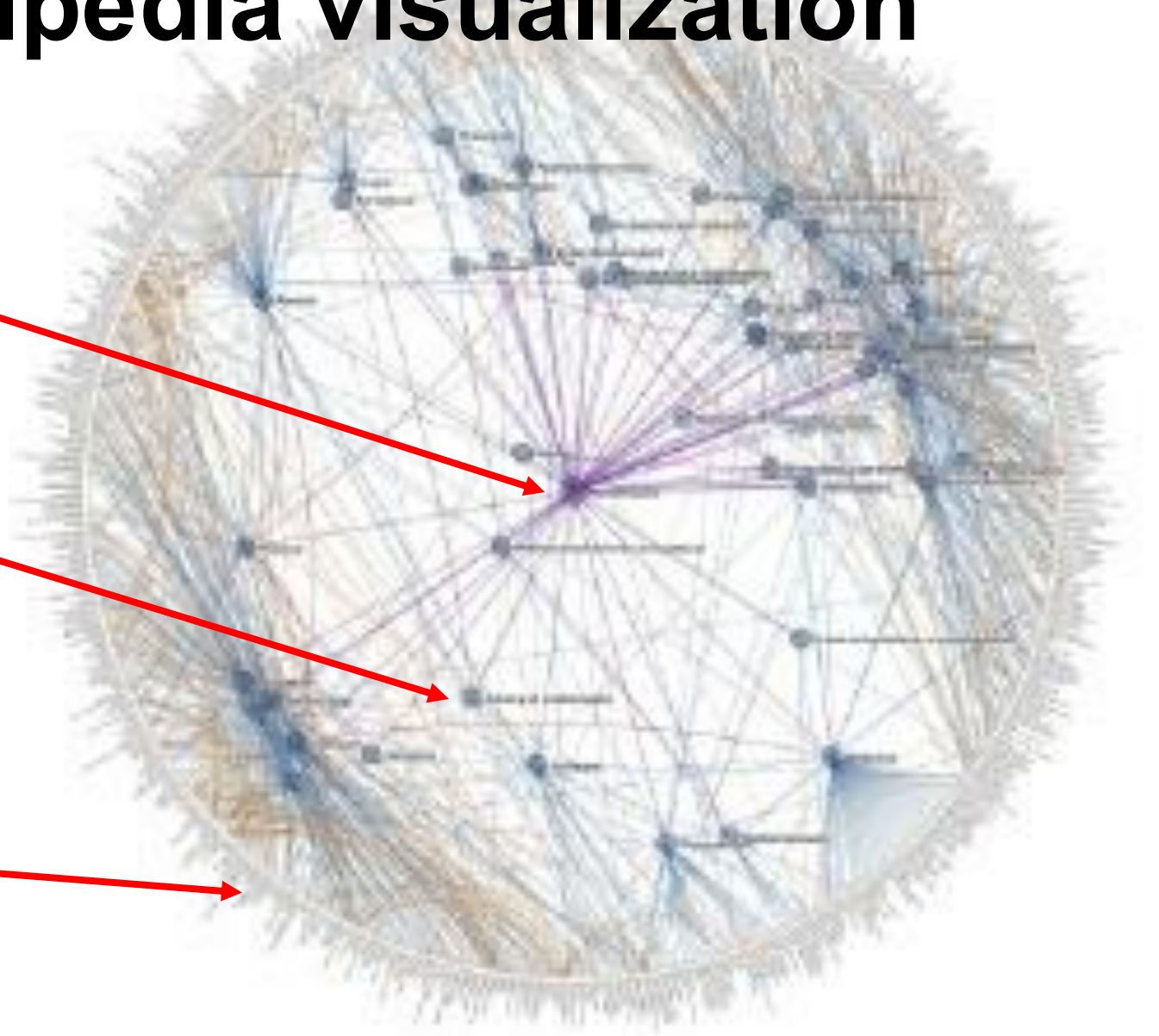
askwiki engine

article search

Done

# Wikipedia visualization

ClusterBall Viz



Mathematics

Nodes inside  
ball one hop  
away

Nodes on ball  
edge are 2  
hops away

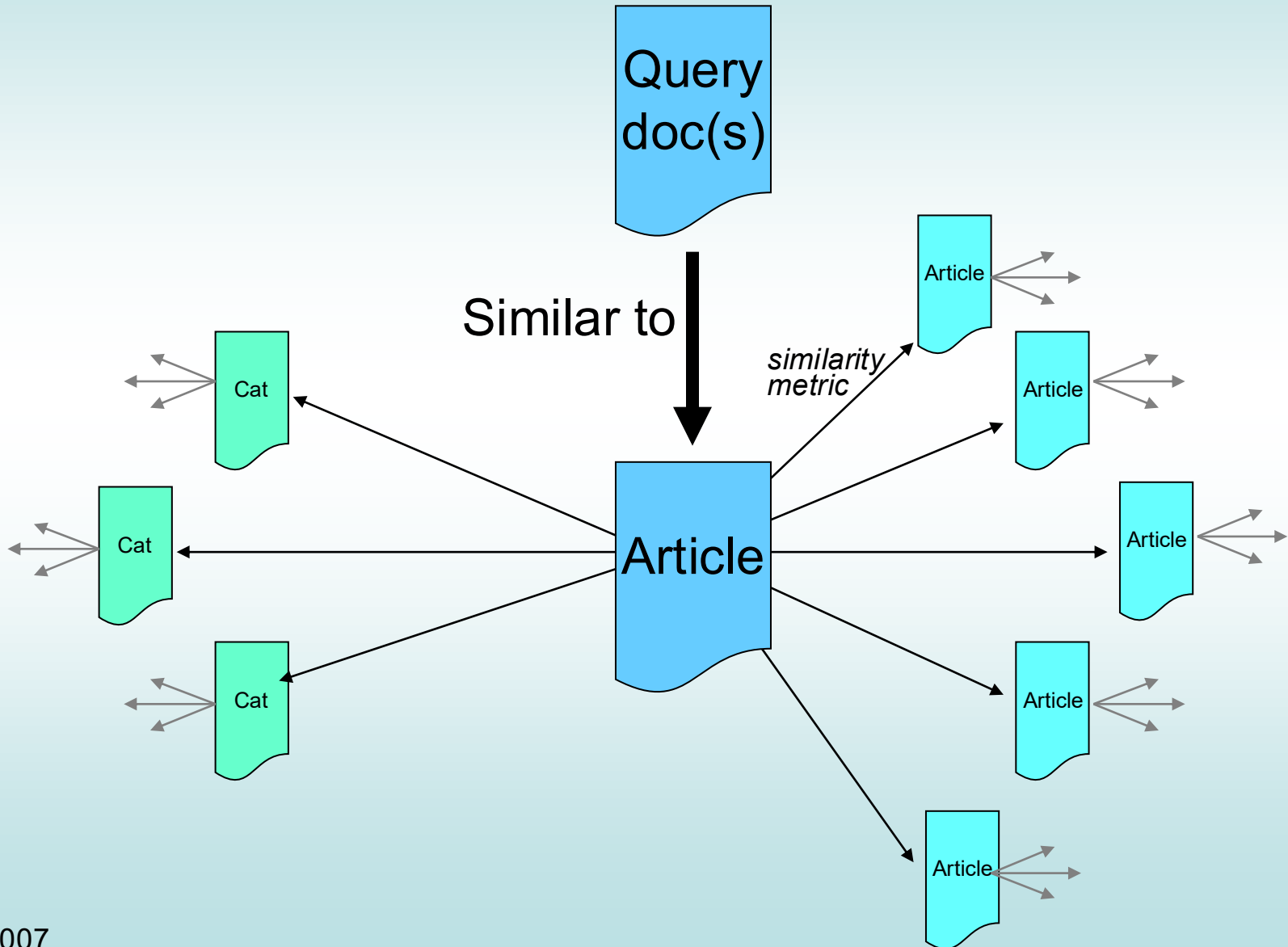
# Preparing the data

- Download Nov 2006 Wikipedia article XML dump (13G)
- Index the ~2.6M articles in Lucene IR system
- Extract **article** and **category** graphs, put in DB
  - ~ 180K categories, 375K category links
  - ~ 90M article-article links
- Cleanup index and graphs by removing administrative & “junk” pages/categories
  - “Articles needing references”
  - “1998”

# Experiments

- Goal: given one or more documents, compute a ranked list of the top N Wikipedia **articles** and/or **categories** that describe it.
- We've explored many ideas to improve accuracy, not unlike designing a light bulb
- Basic metric: document similarity between Wikipedia article and document(s)
- Variations: role of categories, eliminating uninteresting articles, use of spreading activation, using similarity scores, weighing links, number of spreading activation pulses, individual or set of query documents, etc, etc.

# Key Structures



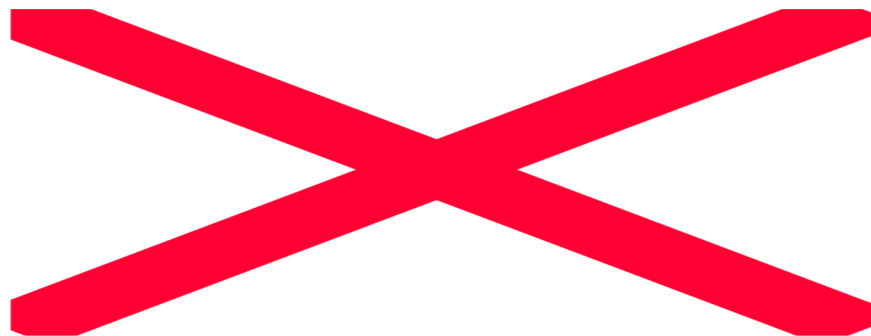
# Experiments

- (1) Rank categories associated with N most similar articles by their frequency
- (2) Like (1) but weight categories by document similarity
- (3) Like (1) but use spreading activation in category graph to elect best categories
- (4) Find top N articles, use spreading activation in article graph (after removing *weak links*) to find best articles

# Evaluation

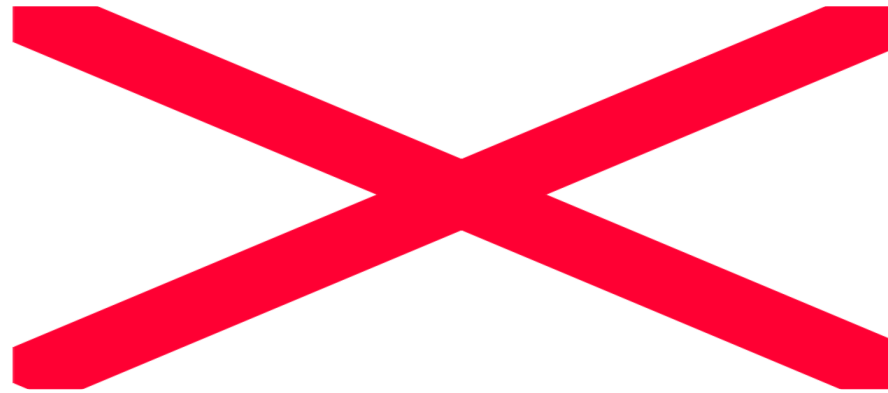
- An initial informal evaluation compared results against our own judgments
- Used to select promising combinations of ideas and parameter settings
- Formal evaluation:
  - Select 100 Wikipedia articles for testing; remove from Lucene index and graphs
  - For each, use methods to predict categories and linked articles
  - Compare results using precision and recall to known categories and linked articles

# Category prediction evaluation



- Spreading activation with two pulses worked best
- Only considering articles with similarity  $> 0.5$  was a good threshold

# Article prediction evaluation



- Spreading activation with one pulse worked best
- Only considering articles with similarity  $> 0.5$  was a good threshold

# Next Steps

- Systematically explore feature combinations/parameters using ML techniques
- Construct a Web-based API and demo system to facility experimentation
- Add Wikitology terms to documents & queries in an IR system to improve performance
  - Using TREC 8 data & JHU/APL Haircut
- Cross-doc entity co-reference for HLTCOE
- Exploit parallel execution on cluster

# Conclusion

- Our initial experiments showed that the Wikitology idea has merit
- Wikipedia is increasingly being used as a knowledge source of choice
- Easily extendable to other wikis and collaborative KBs, e.g., Intellipedia
- Computationally feasible with spreading activation taking the most time
- We are still working to refine the technique