# Freebase: A Shared Database of Structured General Human Knowledge

**Kurt Bollacker, Robert Cook, Patrick Tufts**
Metaweb Technologies, Inc.
631 Howard St.
San Francisco, California 94105

## Abstract

Freebase is a practical, scalable, graph-shaped database of structured general human knowledge, inspired by Semantic Web research and collaborative data communities such as the Wikipedia. Freebase allows public read and write access through an HTTP-based graph-query API for research, the creation and maintenance of structured data, and application building. Access is free and all data in Freebase has a very open (e.g. Creative Commons, GFDL) license.

## Introduction

Over the past two decades, there has been a grand movement of information previously stored and used in isolated systems and repositories toward the public commons that is today's Internet. The popular adoption of the World Wide Web (Web) by large numbers of users has encouraged the growth, diversity, and popularity of (often specialized) public data repositories. Examples include those for visual media (Flickr 2007; YouTube 2007), encyclopedic knowledge (Wikipedia 2007), Scientific Publications (Bollacker, Lawrence, & Giles 1998) and software (SourceForge 2007).

While the Web is an excellent system for the presentation and access of publicly available information, it lacks intrinsic features for organizing this information around the semantics of information item content and/or metadata. Thus, simple keyword-based text indexing is the primary search tool for most Web users. The Semantic Web (Berners-Lee, Hendler, & Lassila 2001) is an approach toward adding rich, deep, structured semantics to Web pages and information repositories. It has evolved to include a set of standard formats, protocols, and organizational conventions that allow the creation of ontological structures and the organization of Web-based information using those structures. The hope of the Semantic Web is that "meaning" rather than keywords can be used to index and organize information. While there has been substantial progress in Semantic Web research and implementation, stable, large-scale public implementations of semantic indexing technologies are still rare. Implementation examples include graph stores suitable for automated reasoning (Lenat 1995), semantic organization of encyclopedic knowledge (Krotzsch, Vrandecic, & Volkel 2005), and large-scale structured information extraction efforts (Auer & Lehmann 2007).

## Freebase

Efforts towards the realization of the Semantic Web's vision have been impeded by a lack of concentrated resources necessary to build an infrastructure large and dense enough to make it immediately useful, especially to laymen beyond the research and technical communities. In the spirit of the Semantic Web and the Wikipedia, we have created **Freebase**, a collaboratively built, graph-shaped database of structured general human knowledge. It is capable of supporting a great diversity of structured data with good performance and able to scale to large numbers of users. Rather than simply a proof-of-concept research project, a substantial effort has been made to make Freebase a stable, practical platform for the wide collaborative creation, organization, and research of the world's public structured information. It shares many characteristics (and is very compatible) with several existing Semantic Web efforts, but emphasizes practical implementation over architectural elegance or perfect standards adherence. In particular, Freebase's features have been designed to promote its use as a public "Data Commons" to be used by any researcher, contributor, or data seeker who sees fit to use it.

### Components Of Freebase

Technically, Freebase consists of the following components:

- **A Graph-Shaped Data Store:** This is a scalable, tuple store with some built-in query planning and optimization capabilities. Similar in flavor to the reversion tools in many wiki systems, the Freebase data store supports complete "undo" of every edit since the database's first insertion, through an integrated versioning mechanism.
- **A Large Data Object Store (LOB):** This is a store of large data objects such as text documents, images, sound files, and software. LOB objects are indexed and annotated in the graph store.
- **A Public HTTP-Based API:** The primary method of access to Freebase is through its public HTTP-based API. Queries and answers are formulated using Javascript Object Notation (JSON) syntax. A novel feature of this API is that writes as well as reads are possible

through this API. We have developed a graph-query language known as the "Metaweb Query Language" (MQL) for access through the Freebase API. MQL is designed for ease of use, write access, and scalability on a *shared* graph store, as compared to more expressive (but read-only) traditional graph query languages (e.g. SPARQL,SeRQL,RQL). Because of this difference, MQL is well suited to data mining, search, graph analysis, and information retrieval applications over large, widely shared, continuously updated data sets. Important MQL features include graph-shaped structural data matching mixed with approximate string matching of literals, cursors for scalable retrieval of search result sets, automatic fine-grained attribution of all data, and intrinsic versioning to allow search over obsolete and deleted data items.

- **A Lightweight Typing System**: Rather than a system of rigid ontologies, Freebase contains tools for the collaborative design of simple types and properties. There is no intrinsic hierarchy of categories or canonical world view of all knowledge. Conflicting and contradictory types and properties may exist simultaneously in order to reflect users differing opinions and understanding.
- **An Easy-To-Use Web UI:** Casual and non-technical users can use Freebase's Web UI to search, browse, create, and edit the data in Freebase at a human scale. Novel features include interactive auto-reconciliation of topics and type-based filtering of search results.
- **A Substantial Initial Data Set:** An emphasis has been placed on on the early seeding of Freebase with data sets of interest to the general population, rather than those that are highly esoteric and specialized. This hopefully results in greater heterogeneity of structure and content, that is more representative of the world's sum of general knowledge. The current data in Freebase consists of millions of concepts (topics) and tens of millions of relationships between those topics. Areas of initial seeding include popular culture (films, music, books, sports), location information (restaurants, geolocations, businesses), scientific information (biological taxonomic and genetic information, astronomy), and general knowledge (Wikipedia). While this data is already useful, we are making efforts for it to grow quickly over time in both quantity and connectedness.
- **A Creative Commons License:** By default and with few exceptions, all users who contribute to Freebase agree to a *Creative Commons Attribution* (CC-by) license (Berry & Moss 2005) for their data.

## Freebase Users

While we expect a great diversity of applications and users of Freebase over time, initially, we are building features to support the following classes of users:

- **Data Contributors:** These users are holders of data that would benefit from placing their data into and providing public accessing from a structured, collaboratively edited, graph-based store.
- **Application Builders:** Those who are interested in building public data services that access the data in Freebase will be supported through the API and provided a variety of demonstration source code examples.
- **Researchers:** Researchers in areas such as entity extraction and reconciliation, data mining, the Semantic Web, information retrieval, ontology creation and analysis, and graph analysis can use the Freebase API to support their work.

## Demonstration

For the AAAI 2007 Intelligent Systems Demonstration, we will show the Freebase API and Web UI and their use in research applications such as data mining, search, information retrieval, and other research applications, and encourage research usage of this public service.

As of April 2007, Freebase is providing an "alpha" level public service for users. We expect that within a small number of months, we will expand to a "beta" service that allows large scale anonymous read access and an increased number of users overall. We encourage users to request an account at www.freebase.com or send e-mail to alpha@metaweb.com.

## References

Auer, S., and Lehmann, J. 2007. What have innsbruck and leipzig in common? extracting semantics from wiki content. In *Proceedings of the European Semantic Web Conference (TO BE PUBLISHED)*.

Berners-Lee, T.; Hendler, J.; and Lassila, O. 2001. The semantic web. *Scientific American* 284:34–43.

Berry, D. M., and Moss, G. 2005. The libre culture manifesto.

Bollacker, K. D.; Lawrence, S.; and Giles, C. L. 1998. CiteSeer: An autonomous web agent for automatic retrieval andidentific ation of interesting publications. In *Proceedings of the Second International Conference on Autonomous Age nts*, 116–123.

Flickr. 2007. The flickr image sharing service. [on the internet. `http://www.flickr.com`].

Krotzsch, M.; Vrandecic, D.; and Volkel, M. 2005. Wikipedia and the semantic web - the missing links. In *Proceedings of Wikimania 2005, Frankfurt, Germany*.

Lenat, D. B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(11):33–38.

SourceForge. 2007. The sourceforge software hosting service. [on the internet. `http://www.sourceforge.net`].

Wikipedia. 2007. Wikipedia: The free encyclopedia. [on the internet. `http://www.wikipedia.org`].

YouTube. 2007. The youtube video sharing service. [on the internet. `http://www.youtube.com`].