

# Alternatives to Author-centric Knowledge Organization

Rama C. Hoetzlein

Media Arts & Technology Program, University of California Santa Barbara, Santa  
Barbara, California, USA

[rch@umail.ucsb.edu](mailto:rch@umail.ucsb.edu)

**Abstract.** The digital era of the document begins with the hyper-textural, yet still author-centric model of HTML. Since then, we find both the migration of large-scale traditional document collections to online resources (Siemens 2006), and the development of community-driven resources such as Wikipedia. In these secondary resources, the authorship of meta-data and factual reference knowledge is extended to a larger, global community. However, overlapping opinion often leads to issues of authoritative control (Viegas 2004). While the traditional document is typically an individual, or small-scale collaboration, the social aspect of the document becomes much more critical in global online systems. Different strategies are needed to maintain or track authorship in documents, or knowledge resources, with thousands of authors.

Quanta is presented as an alternative to author-centric documents using a custom built, non-relational database of sentences. While encyclopedias have traditionally represented facts as a body of words in an article, Quanta represents knowledge on an atomic level, explicitly maintaining each word. This granularity introduces new ways to represent and interact with a text, or a collection of facts, and presents the possibility of assigning authorship to individual sentences.

By taking the sentence as the unit of knowledge, and representing it in machine-readable form, written language is connected to the database in ways similar to semantic networks (extended in this work with hypergraphs). All words in a sentence are hyperlinked to the text, and to larger bodies of knowledge. As Quanta makes several assumptions about language, it may be viewed as a supplement to the document, yet at a much deeper level than meta-data. By introducing filtering and visualization tools, an atomic description of language allows for overlapping, related, and even inconsistent, distinctly authored facts to co-exist in a single system.

**Keywords:** Quanta, knowledge environment, representation of text, authorship, information studies, encyclopedias, language-based systems

## Introduction

The Encyclopédie, printed and published in French by Andre Francois Le Breton in 1751, was one of the first large scale reference works. Containing 35 volumes, and over 71,818 articles, with over 120 contributors listed, the Encyclopédie was "vastly greater in scale than any [earlier] English works" (Lough 1971). The primary author, Denis Diderot, originally estimated completion in 1754, but the work was not completed until 1772. The central reason for this may be attributed to the immense scope and vision of the work:

"The work whose first volume we are presenting today has two aims. As an Encyclopedia, it is to set forth as well as possible the order and

connection of the parts of human knowledge. As a Reasoned Dictionary of the Sciences, Arts, and Trades, it is to contain the general principles that form the basis of each science and each art, liberal or mechanical, and the most essential facts that make up the body and substance of each." (Schwab 2009)

Diderot recognized that such a work could not be completed by one man, but that it must involve a "society of men of letters and skilled workmen, each working separately on his own part." The distinction of articles according to fields of study certainly helped the large scale of the collaboration, and is a primary feature of modern encyclopedias (e.g. Encyclopedia Britannica). However, the writing of the Encyclopédie was not without major challenges. Only 20 of the 120 contributors were paid. In a few cases, paid authors fled or produced little work. One author was condemned due to external events. Many authors were paid much less than the amount of work they produced (Lough 1971). Collaborative authorship always presents a range of problems, mostly more social and human than practical. Nonetheless, notable persons from many distinct fields of study were brought together to complete the work.

Encyclopedias are an example of *collaborative authorship*, in which the authors generally know one another and have agreed to a strategy for publication and division of labor. Subramanyam reviews the types of collaboration, and finds that collaboration affects the visibility and productivity of modern researchers (Subramanyam 1983). Harande finds that in certain fields, such as technology, productivity is not necessarily linked to collaborative authorship (Harande 2001). We can distinguish this from *collective authorship*, in which the authors are anonymous, unfamiliar with one another, or have not decided on delineations of labor. Nearly all historic authorship is of the first type, including most reference works, encyclopedias, dictionaries and scientific papers. The Internet is the first unique example of collective authorship. Considered as a whole, it represents a semi-anonymous, collectively authored, delimited text. The introduction of HTML permits hyper-linked texts in the decentralized space of the web (Berners-Lee 1999). However, the fact that individual texts are personally owned and maintained causes a delineation of web pages along boundaries of authorship. Thus, publication on the Internet is similar to the publication of an individually authored text. The Internet acts more as a repository for these texts, which is further supported by the fact that the Internet is not a reference work, or condensation of knowledge, but is usually "searched" for sources.

Collectively authored online encyclopedias, such as Wikipedia, represent a different kind of text. Wikipedia was conceived of as a free online encyclopedia, yet the writing of content was extremely slow as it initially followed traditional models of authorship. Jimmy Wales and Larry Sanger realized that a text based on a *wiki*, software for online content creation, would allow many anonymous authors to contribute to the same text. Wikipedia now has over 2.9 million articles, with an average of 20 edits per page, while frequently edited articles can have over 2000 edits. Since the average article length does not change significantly,<sup>1</sup> this means that each paragraph or sentence is contributed by a different anonymous author. The text is highly granular in its authorship.

While the scale of Wikipedia is immense in comparison to other references, it has also been heavily criticized. The number of administrators capable of resolving conflicts or locking pages is only 1675, far fewer than the number of

---

<sup>1</sup> All facts from the Wikipedia History page ([http://en.wikipedia.org/wiki/History\\_of\\_Wikipedia](http://en.wikipedia.org/wiki/History_of_Wikipedia)). Wikipedia states the average article length is a little over half that of an Encyclopaedia Britannica article.

contributors. In addition, administrators have particular areas of interest but may not be authorities in these fields. A unique phenomenon of Wikipedia is the "revert war", in which two or more authors competitively modify the same content. Viegas et al. find that revert wars are not limited to controversial topics and may take many different forms (Viegas 2004). As they mention, the "neutral point of view" philosophy is Wikipedia's proposed solution, with discussions taking place in a separate space. Contributors may be required to provide citations for each statement, which is often impossible in fields such as the humanities where discussion is more interpretive. One cannot express an individual opinion on a literary work, for example, unless that opinion is backed by a previous "cited" author. In some cases, it was found (by the author) that citations had nothing to do with the statement, and were provided merely to lend authority. Anonymous, collective authorship thus presents several unique challenges to the creation of reference works.

### **Knowledge Organization**

An alternate approach to the collective authorship of tertiary texts may be motivated by the fields of library science and knowledge organization, which extend primarily from the concept of the database. A definition of *knowledge organization*, provided by the librarian Berwick Sayers, reveals a close relationship to the card catalog:

"[knowledge organization is] not only the general grouping of things for location or identification purposes; it is also their arrangement in some sort of logical order so that the relationship of the things may be ascertained." (Sayers 1959)

In knowledge organization greater importance is placed on the *arrangement* rather than *collation* of the primary sources, as with encyclopedias. There may be many books on the topic of physics present in a card catalog, with different views on a particular theory, yet the purpose of the catalog is not to resolve or summarize different physicists' views, but simply to maintain each interpretation as a distinct text. This approach is in contrast with encyclopedias, which attempt to condense knowledge from many sources.

There are several benefits to data-centric knowledge organization. A card catalog is trivial to collectively author, since there is a one-to-one relationship between a catalog entry and a primary source. A librarian may easily enter a new book without being concerned that the content may overlap or conflict with another similar book. This is also their drawback, however, since a card catalog provides many references but no detailed information regarding the field.

Unlike encyclopedias, a card catalog does not attempt to communicate a summary of human knowledge, it only attempts to organize it. This simplifies authorship, while the drawback is that one must still ultimately browse many primary sources to find meaning. An encyclopedia presents meaningful content, but introduces conflicts of authorship due to the need to simplify and merge original sources which may contain different views. Ideally, we would like a system for human knowledge which balances organization with levels of meaning.

Vannevar Bush, with the hypothetical Memex, describes the ideal features of a collectively authored, centralized knowledge system.<sup>2</sup>

---

<sup>2</sup> Centralized in the sense of collected knowledge (bringing together), while the underlying storage may be decentralized in the physical sense and/or also decentralized in the managerial sense.

The owner of the memex, let us say, is interested in the origin and properties of the bow and arrow. He has dozens of possibly pertinent books and articles in his memex. First he runs through an article, finds an interesting but sketchy article, and leaves it projected. Next, in a history, he finds another pertinent item, and ties the two together. (Bush 1945)

Interestingly, this description covers libraries (primary sources), encyclopedias (condensations), and timelines (systematic organization) - different scales of authorship and organization present in a single system. The realization is that human thought is at times broad, in need of summary overviews, yet also specific once details are grasped.

The implementation of the Memex, written in 1945, is described as a system of "books of all sorts, pictures, periodicals and newspapers" placed on microfilm, and fitting into a space the size of a desk, so that "the matter of bulk is well taken care of by microfilm". A wonderful idea until one realizes that the US Library of Congress would require a full city block of microfilm itself (as it does), and the entire internet would need sixteen city blocks.<sup>3</sup> With modern storage, however, this is not the primary issue. More challenging, the means to organize, navigate, collect and summarize this knowledge presents many theoretical and practical problems, primary among them that each reader, or organizer, will interpret the material differently.

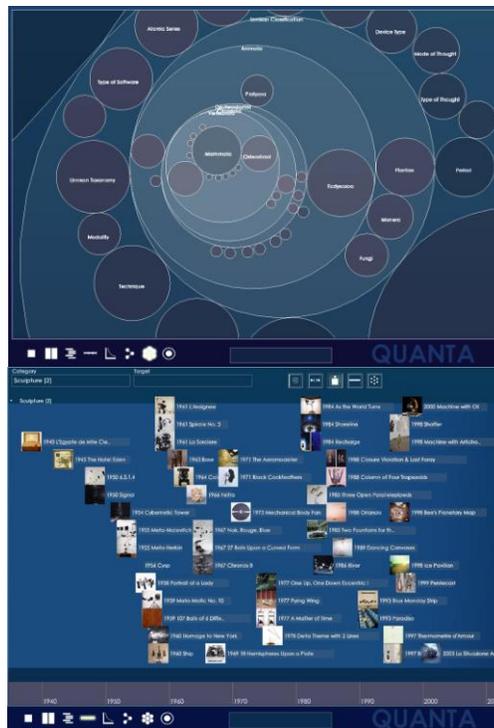


Figure 1. Screenshots of Quanta showing a) a timeline filtered to show contemporary sculpture, and b) a circle-packing view of the Linnean ontology for animals.

<sup>3</sup> One roll of Microfilm = 0.002 terabytes. Largest electronic storage (2008) = 1000 terabytes. Library of Congress = 25,000 terabytes. Entire Internet (est) = 400,000 terabytes. (Src: Intl. Data Corporation)

### **Quanta: A system for Language-based Knowledge Organization**

At present, information systems, card catalogs, and journal indices are almost all stored in relational databases. The relational database model became the industry standard in the 1970s largely due to its simplicity and relative ease of use at the time, while the competing model of network databases continued only as research. However, it is well known that relational databases have critical restrictions on their grammatical expressiveness and flexibility (Levene 1998). Thus, while modern encyclopedias struggle with being too unstructured, leading to tedious organization and conflicts in authorship, knowledge systems tend to be too structured, allowing for only indexical information rather than deep facts.

Quanta is an experimental system designed to be both structured and also grammatically rich, thus residing at a middle point between encyclopedia (language) and databases (objects). Facts are represented in Quanta by taking the sentence as a unit of knowledge,<sup>4</sup> and the word as its atomic component, so that the language of Quanta consists of phrases with word and phrases separated by vertical bars, such as:

```
light | has | constant | speed | in | vacuum | CITE | Einstein | ENTRY | J.
Smith
life | exists | on | mars | BELIEF | J. Smith | RATING | 2
Hamlet | was | written by | Shakespeare | ENTRY | J. Smith
J. Smith | is an | active user
```

This structure is similar to the Prolog programming language, based on prepositional logic. Note that the capitalized words delimit second order phrases regarding the statement. The unique aspect of Quanta, however, is that these sentences are embedded in a hypergraph database, providing relational context in addition to grammar (Hoetzlein 2007). By representing language in this way, it is possible to express complex ideas while at the same time automatically linking these ideas to every other context in which they appear. A search for the word 'fish,' for example, is automatically connected via the hypergraph to every context in which fish appears - types of fish, clouds that look like fish, robotic fish, cooking fish - all appear when the idea of fish is explored.

### **Authority versus Filtering**

It is interesting to observe that in many Wikipedia articles, the practice of placing references after each sentence is becoming increasingly common. Conflicts between ideas must be continually resolved through careful editing or administrative intervention to express both perspectives while maintaining readability. The authority of administrators has led to a number of criticisms as individuals with strong biases may be frequent editors of a particular topic. The task of determining which ideas should be present is a challenging one. Does an implausible theory by a lesser physicist deserve to be present on a page with well-known theorists? Should personal descriptions of God be allowed on pages about religions? What if they are by notable saints? Should views that man never landed on the moon be allowed on the page for space flight?

The syntactic level of Quanta provides a format which can distinguish between the writer (author) of an idea and the source (citation) of that idea. This helps to contextualize the fact relative to the source of the information, but also according to the biases of the person who offered the fact in the first place. In an online system, in which participants log in, the author can be automatically associated with each idea which is entered.

---

<sup>4</sup> Sentences have variable lengths and structure whereas relational records are fixed length with fixed fields.

At a higher level, a novel approach to collective authorship of encyclopedia would enable all points of view to co-exist. The primary reason for a "revert war" is that a biased, anonymous author has the power to delete the ideas of another individual. However, this situation does not occur in reality. No one can erase our beliefs, as much as they might try to through persuasion or argument. Similarly, the ideal knowledge system would allow all perspectives to co-exist simultaneously so that anyone may add an authoritative fact, or opinion, regardless of their correctness. Such a system may be especially well suited to children, allowing perspectives from all ages to co-exist along side authoritative sources on a particular topic. Filtering allows a grade school classroom to show similar ideas from the same age range, or the researcher to filter only for key sources.

A collective, *additive-only* text presents several challenges. The first is how the reader distinguishes authoritative facts from incorrect ones. As sentences are atomic, Quanta automatically assigns authorship to every fact entered. Thus, it becomes possible to filter a page based on all facts from a particular authoritative individual or group. To see factual knowledge on astronomy, for example, one could filter based on authors with degrees in this field, or by specific authors. Finally, the system could allow readers to rate the accuracy of any statement, so that collective, statistical agreement on the correctness of ideas may still be recorded. The need for content administration no longer exists, as all authorship is individual and additive.

Another challenge presented by collective authorship is vandalism. However, we notice that acts of vandalism due to deletion are eliminated since deletion is not allowed, and the meaning of vandalism is diminished since each fact is automatically tagged with the author at the time of entry. Since Quanta maintains author associations internally, to control vandalism simply involves filtering by these authors. Anonymous log in and authorship is also permitted, with facts tagged as anonymous, while the reader may just as easily view anonymous facts as filter them out.

Providing fine-grained filtering for readers resolves many issues, and more closely mimics our human perception of reality. When we encounter views which go against our beliefs we may filter them out completely (Wason 2004). This is due to the basic necessity of simplifying our experiences. People attempt to live in places which conform to their system of beliefs. Rather than restrict authorship, which results in power issues, precise filtering tools places the control and bias entirely at the reader's discretion (i.e. choosing where to go).

The perspective of authorship presented here is that there are no incorrect views, nor completely correct ones, only degrees of authority. In this sense, the text is comparable to a library. No one may remove books, but anyone may add one or check one out. Similarly, in this experimental system, no one writing a text may remove facts provided by others, but anyone may add new facts.<sup>5</sup> This strategy of (non-) authority is not maintained by a select group of individuals, but through an automated and systematic process of assigning authorship at the time facts are created, and allowing readers precise control over filtering of texts. While very similar in spirit and motivation to open authorship on the Internet, the design presented is for a system which is logically structured to provide deep organization for general types of information, and to carefully track authorship to simplify navigation and filtering.

---

<sup>5</sup> It may be argued that unrestricted addition would consume large amounts of space. However, the grammar structure of Quanta allows it to be stored using a dictionary-compression scheme, with words stored as numbers. Thus, the store needs are significantly less than plain text articles.

### **Results and Limitations**

Quanta was designed as a knowledge organization and authoring system to allow different levels of knowledge to co-exist, both systematic (relational) and encyclopedic (grammatic). In practice, this was accomplished with the creation of a custom, non-relational database written in C++, and a larger meta-system for querying, visualizing and navigating that data. Design decisions for the system may be found in the original work on this project (Hoetzlein 2007). The initial results were primarily focus on data storage and representation, resulting in an offline prototype database.

As the data representation of Quanta is both semantic and systematic, consisting of a rich grammar and inter-connected database, it becomes possible to create novel visualizations of content. Several visualizations developed include timelines of arbitrary concepts (zoomable on both axes, time and detail), scientific graphs of any two numeric values, circle packing as a two-dimensional mapping metaphor for conceptual ideas, and ontology trees to view individual belief hierarchies (Figure 1). In each mode, filtering allows different sets of concepts to be viewed. These real-time visualizations demonstrate that facts can be efficiently represented and queried in the structures described here.

Recent work has focused on an online system to introduce the social aspects of additive authorship. The core database has been redesigned to provide the ability to track authorship on individual sentences, along with other types of self-referential information. A working online system is still in development, while the offline prototype demonstrates many of the concepts described here with examples from the fields of computer graphics, painting, philosophy, mineralogy and chemistry.

A related project is the PReE/ProSE system for social computing, a collaboration between the Social Computing Group at the University of California Santa Barbara and the INKE project at the University of Victoria, which introduces the notion of professional readership as an extended collection of historic persons, contemporary critics, authors, readers and the associated primary sources surrounding them.

Each of these systems has benefits and limitations. Quanta uses a custom database, and is thus similar to OpenCyc, a database of common sense maintained in Lisp format. Both systems require significant low-level development. While the scope of Quanta is both deep and broad, focusing on all of human knowledge, an online system allowing sentence-level tracking of authorship as described is still in development. Wikipedia is based on a content authoring system, and is more closely related to document authorship than to databases. PReE/ProSE uses a relational MySQL database, and while currently available, its granularity is not sufficient to permit the range of general knowledge found in Quanta. Despite these differences, each system described makes unique advances in authorship, readership and organization according to their goals.

### **Conclusions**

With the exception of libraries, in every knowledge organization system available - from print encyclopedia to online sources - the concept of authority restricts knowledge according to the biases of the administrators of that content. This is also true of libraries, as any library must select which materials to include. The internet itself is the only modern system which allows unlimited addition of knowledge without authoritative control. Yet, the internet is non-encyclopedic (uncondensed). As suggested here, it may possible to design future knowledge organization systems which remove central control by providing unlimited additions, like the Internet, but with fine-grained (word level) tracking and filtering tools for authors and readers.

The anonymous, collaborative authorship of large references is a problem unique to the digital humanities as the challenge of organizing global knowledge is likely to require on-going human intervention in the areas of information science, computation and literature, philosophy and others. An alternative to current author-centric knowledge systems is presented here, while issues related to the study of control, authorship, and reading practices of digital reference works remain open areas for further research.

## Works Cited

- Berners-Lee, Tim. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. New York: Harper Collins, 1999. Print.
- Bush, Vannevar. "As We May Think." *The New Media Reader*. Ed. Noah Wardrip-Fruin and Nick Montfort. Cambridge: MIT, 2003. Print.
- Harande, Y.I.. "Author Productivity and Collaboration: An Investigation of the Relationship Using the Literature of Technology." *Libri* 51 (2001): 124-27. Print.
- Hoetzlein, Rama. *The Organization of Human Knowledge: Systems for Interdisciplinary Research*. Masters thesis. 2007. Print and Web. <<http://www.rchoetzlein.com/quanta>>.
- Levene, Mark. "On the Information Content of Semi-Structured Databases." *Acta Cybernetica*\_13.3 (1998): 257-75. Print.
- Lough, John. *The Encyclopédie*. New York: David McKay, 1971. Print.
- Sayers, W.C. Berwick. *A Manual of Classification for Librarians and Bibliographers*. London: Andre Deutsch, 1959. 5. Print.
- Schwab, Richard N. "Preliminary Discourse to the Encyclopedia." *The Encyclopedia of Diderot & d'Alembert: Collaborative Translation Project*. Web. 15 Aug 2009. <<http://quod.lib.umich.edu/d/did>>.
- Siemens, Ray, Eric Haswell, Gerry Watson, Alastair McColl and Karin Armstong. "Integrating Tools into Professional Academic Processes: A First Look at the Renaissance English Knowledgebase (REKn)." *Bringing Text Alive: The Future of Scholarship, Pedagogy, and Electronic Publication*. University of Michigan. Rackham Graduate School, Ann Arbor, MI. 14-17 Sept. 2006. Address.
- Subramanyam, K. "Bibliometric Studies of Research Collaboration: A review." *Journal of Information Science* 6.1 (1983): 33-38. Print.
- Viegas, Fernanda, Wattenberg, M. and Kushal, D. "Studying Cooperation and Conflict between Authors with history flow Visualizations." *ACM SIGCHI Special Interest Group in Computer-Human Interaction*. Vienna Austria. 2004. Address.
- Wason, Peter Cathcart. "Confirmation Bias." *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Ed. Margit E. Oswald and Stefan Grosjean. Hove: Psychology, 2004. 79-96. Print.