In a prior post[^1] I argued that the recent coverage of Wikipedia has focused too much on one aspect of the online reference source’s openness—the ability of anyone to edit any article—and not enough on another aspect of Wikipedia’s openness—the ability of anyone to download or copy the entire contents of its database and use it in virtually any way they want (with some commercial exceptions). I speculated that, as I discovered in my data-mining work with H-Bot[^2], which uses Wikipedia in its algorithms, having an open and free resource such as this could be very important for future digital research—e.g., finding all of the documents about the first President Bush in a giant, untagged corpus on the American presidency. For a piece I’m writing for D-Lib Magazine[^3], I decided to test this theory by pulling out significant keywords and phrases from matching articles in Wikipedia and the Encyclopaedia Britannica on George H. W. Bush to see if one was better than the other for this purpose. Which resource is better? Here are the unedited term lists, derived by running plain text versions of each article through Yahoo’s Term Extraction web service. Vote on which one you think is a better profile, and I’ll reveal which list belongs to which reference work later this week.

**Article #1**

president bush
saddam hussein
fall of the berlin wall
tiananmen square
thanksgiving day
american troops
manuel noriega
halabja
invasion of panama
gulf war
help
saudi arabia
united nations
berlin wall

**Article #2**
president george bush
george bush
mikhail gorbachev
soviet union
collapse
reunification of germany
thurgood marshall
union
clarence thomas
joint chiefs of staff
cold war
manuel antonio noriega
iraq
gorge
nonaggression pact
david h souter
antonio noriega
president george

This entry was posted on Monday, January 30th, 2006 at 12:07 pm and is filed under APIs[^4], Text Mining[^5], Web Services[^6], Wikis[^7], Yahoo[^8]. You can follow any responses to this entry through the RSS 2.0[^9] feed. You can [leave a response][10], or [trackback][11] from your own site.

**References**

1. ^ a prior post (www.dancohen.org)
2. H-Bot (chnm.gmu.edu)
3. D-Lib Magazine (dlib.org)
4. View all posts in APIs (www.dancohen.org)
5. View all posts in Text Mining (www.dancohen.org)
6. View all posts in Web Services (www.dancohen.org)
7. View all posts in Wikis (www.dancohen.org)
8. View all posts in Yahoo (www.dancohen.org)
9. RSS 2.0 (www.dancohen.org)
10. leave a response (www.dancohen.org)
11. trackback (www.dancohen.org)

Excerpted from Dan Cohen’s Digital Humanities Blog » Blog Archive » Wikipedia vs. Encyclopaedia Britannica for Digital Research
http://www.dancohen.org/2006/01/30/wikipedia-vs-encyclopaedia-britannica-for-digital-research/

READABILITY — An Arc90 Laboratory Experiment
http://lab.arc90.com/experiments/readability