A World Map of Knowledge in the Making: Wikipedia's Inter-Language Linkage as a Dependency Explorer of Global Knowledge Accumulation

T Petzold [a,c], HT Liao [b], J Hartley [a] and J Potts [a,d]; [a] Centre of Excellence for Creative Industries and Innovation, Brisbane, Australia, [b] Oxford Internet Institute, Oxford, UK, [c] Social Science Research Center Berlin, Germany, [d] University of Queensland, Brisbane. Corresponding author: <tm.petzold@gmail.com>.

See <www.mitpressjournals.org/toc/leon/45/3> for supplemental files associated with this issue.

Submitted: 30 September 2011

Abstract

Analysis of Wikipedia's inter-language links provides insight into a new mechanism of knowledge sharing and linking worldwide.

Wikipedia users write and edit articles in more than 260 languages; manual labor and automatic "bots" have together created links across different language versions for articles on the same or similar topics [1]. These links, also known as inter-language links, in effect define what count as "nearly equivalent or exactly equivalent" concepts across languages. The idea of equivalent concepts across a selection of languages, now digitally embodied in the form of more than a quarter-billion inter-language links [2], can provide vital indications of both the knowledge relationship among languages and the currency of concepts among languages (some universal/cosmopolitan and some particular/regional).

During the process of knowledge accumulation, each language develops its own set of inter-language links with other language versions, the totality of which creates a dynamic network of relations amongst the different language versions in Wikipedia. Although basic indicators such as the numbers of articles and active editors have shown disparity between knowledge have-more versus knowledge have-less Wikipedia versions [3], identifying patterns in inter-language links further provides an indication of how such disparity may be structured and distributed. Research on inter-language linking across the World Wide Web, for example, finds that each language has its own densely interconnected web that is loosely linked to other languages [4], supporting the hypothesis that the *world-wide* web may be interconnected conditioned by geo-linguistic factors [5,6].

The Rise of Middle Powers?

The overall core-periphery network pattern, with English (en), French (fr), German (de) and Spanish (es) versions at the core, is expected, as they are the early-adopters of the Wikipedia project (see also Eric Zachte's animated growth figures [7]).

Entries such as "Wikipedia", already saturated with dense inter-language links in all language versions, are not as useful as the Chinese entry of "敗犬" (a derogatory term imported from Japan to describe unmarried women over 30) to visualize how certain language versions are more interconnected than others. Thus, more meaningful visualization is achieved by limiting inter-language link data to those entries with fewer links, so that the dependency pattern can be better identified from the inter-language link network, as shown for Chinese (zh) in Fig. 1. For each language node, the outgoing links shown visualize the major linking targets, that is, the links that have more than 7.5% of total out-going external links.

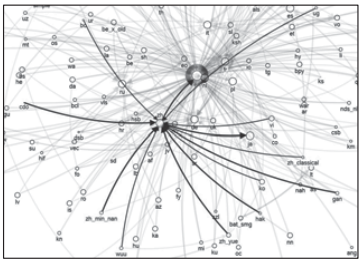

Fig. 1. The Chinese Wikipedia gives more links proportionally to the English and Japanese versions, and receives disproportionally more links from language versions such as Tibetan (bo), Uighur (ug), Vietnamese (vi) and various Chinese dialects, including Cantonese (zh_yue). (© Han-Teng Liao & Thomas Petzold, released under Creative Commons BY-SA license)

Russian (ru), another rising middle power in Wikipedia, contributes more of its links to English and German while receiving – for various historical, cultural and ongoing political economic reasons – from languages such as Azerbaijani (az), Georgian (ka), Latvian (lv), Tajik (tg), Tartar (tt) and others.

The increasing online presence of non-Latin language scripts such as Chinese, Cyrillic or Arabic has not changed the existing Latin-alphabet-only core structure. Yet, we already witness the rise of middle powers of language that accumulate their collection of knowledge and regional influence. While English and other Latin-alphabet languages may remain dominant, especially for scientific and technical knowledge, some languages such as Arabic, Chinese, Hindi and Russian are on the way to reclaim their historical, cultural and even political status and currency in the online environment.

The Next Internet Revolution Won't be in English

While originally adopted mostly by Latin-alphabet users, Wikipedia as a website for knowledge from the world's languages still reflects the core-periphery pattern of knowledge production and exchange online. However, as the Internet infrastructure keeps transitioning to a more multilingual environment [8,9], and the Internet population expands to include more users who use major world languages such as Arabic, Chinese or Hindi, the semi-core and semi-peripheral status of certain languages will play a significant role in the knowledge production and dissemination process, regionally or even globally.

The inter-language dataset of the Wikipedia project is a powerful albeit limited representative of global knowledge accumulation online. However, a more detailed understanding of the core-periphery network patterns of knowledge accumulation and dissemination among different languages (and thus corresponding regions) has raised important issues regarding aggregating/connecting vis-à-vis disaggregating/disconnecting dynamics of knowledge activities worldwide.

References and Notes

***** This paper was presented as a contributed talk at Arts | Humanities | Complex Networks – 2nd Leonardo satellite symposium at NetSci2011. See <http://artshumanities.netsci2011.net>.

1. <http://goo.gl/qlgcO>, accessed 3 May 2011.

2. Calculation based on data extracted by Wikimedia Toolserver, Data Base Query 139, accessed 3 May 2011. We thank Daniel Kinzler and Wikimedia Germany for their support.

3. Instead of using a simplified haves/have-nots framework, we borrow the metaphor of have-more / have-less to better understand the emerging in-between; cf. Cartier, Castells & Qiu (2005), The information have-less, *SCID* 40(2), 9-34.

4. D. Ford & J. Batson (2011), *Languages of the World (Wide Web),* Google Research Blog, <http://goo.gl/Uwpjb>, accessed 7 July 2011.

5. T. Petzold (2010), *36 Million Language Pairs,* Discussion Paper, Berlin Roundtables on Transnationality, Social Science Research Center Berlin, <http://goo.gl/Zk64C>, accessed 16 July 2011.

6. H. Liao & T. Petzold (2010), Geo-linguistic dynamics of the WWW, *Cultural Science* 3(2).

7. E. Zachte, 'Animated growth figures per Wikimedia project', <http://stats.wikimedia.org/wikimedia/animations/growth/index.html>, accessed 23 August 2011.

8. <http://goo.gl/iMbtM>, accessed 16 July 2011.

9. H. Liao (2011), *Needing to Have a Voice*, ISD Working Paper series, Institute for the Study of Diplomacy, Georgetown University.

