Integración cultural -y III - CULTURÓMICA

Pedro R. García Barreno
Catedrático emérito, UCM

[versión ampliada y anotada de la publicada en ***Revista de Occidente***, sept. 2015]

«**Poesía en movimiento**»
Editorial. *Nature*.

Esta revisión surge de la lectura del artículo «*Quantitative analysis of culture using millions of digitized books*» publicado por Jean-Baptiste Michel, Erez Lieberman Aiden et al., y pretende divulgar lo que tales autores denominan reconstruir el esqueleto de una nueva ciencia: la aplicación de *big data* a las humanidades o culturómica.

La tradición cultural, la gran conversación, representa nuestro canon. «Originariamente ─escribe Harold Bloom en *El Canon Occidental*─, el canon significaba la elección de libros por parte de las instituciones de enseñanza y, a pesar de las actuales ideas de multiculturalismo, la auténtica cuestión del canon subsiste todavía: ¿Qué debe intentar leer la persona que todavía desea leer en este momento de la historia? [...] El canon es, sin duda, un patrón de vitalidad, una medida que pretende poner límites a lo inconmensurable». Canon que, en la actualidad, lo contextualizamos como «culturómica»: un producto emergente ─por tanto más complejo y con nuevas propiedades─ de la lexicología computacional que estudia el comportamiento y las tendencias culturales humanas mediante el examen cuantitativo de millones de textos digitalizados ─miles de millones de palabras─ utilizando las técnicas de análisis de megadatos (*big data*). En la oscarizada película *Dead Poets Society*, el carismático profesor de literatura ─John Keating─ apremia a sus atónitos discentes leer la introducción de un texto de poesía: «El valor de un poema está determinado por los parámetros sobre dos ejes: su perfección artística y su importancias». «¡*Excrement*!», exclama el «Capitán»; tal es su opinión a la aproximación matemática. «Un poema se siente, no se mide».

«No es la primera vez que una nueva clase de lente influye en como miramos el mundo», escriben E. Aiden y J-B. Michel a poco de comenzar su libro *Uncharted*. Las gafas hicieron de la optometría un buen negocio. Lentes compuestas hicieron asequible el mundo microscópico y permitieron a Galileo desentrañar el misterio del cosmos. En nuestros días, microscopios y telescopios siguen siendo elementos básicos para el progreso de la ciencia. Una nueva y hasta escasos años impensable tecnología está detrás del conocimiento más reciente en astronomía, física, química o biología. «En 2005, cuando ambos éramos estudiantes pasamos mucho tiempo pensando en cómo los científicos habían hecho posible el progreso científico [...] Los dos llevábamos tiempo interesados en el estudio de la historia [...] ¿Podría alguna clase de microscopio estudiar la cultura humana identificando las pequeñas cosas de las que nunca tuvimos noticia, o que un telescopio acercara los acontecimientos ocurridos siglos atrás? [...] Pero algo de este tipo tendría que ser lo suficientemente *cool* para que Harvard lo aceptara en términos de un PhD». Algo así como la utilización de conceptos derivados de la mecánica cuántica ─efecto túnel─ en el desarrollo del microscopio de efecto túnel (0.1 nm de resolución lateral y 0.01 nm de resolución de profundidad). Coetáneo al planteamiento apuntado tenía lugar la eclosión de la última revolución ─*big data*─ que incide de lleno en las ciencias experimentales o en el comercio. *Big data* crea y almacena el registro histórico de las actividades sociales. «*Big data* cambiará las humanidades, transformará las ciencias sociales y renegociará las relaciones entre comercio y academia», concluyen Erez y Jean-Baptiste.

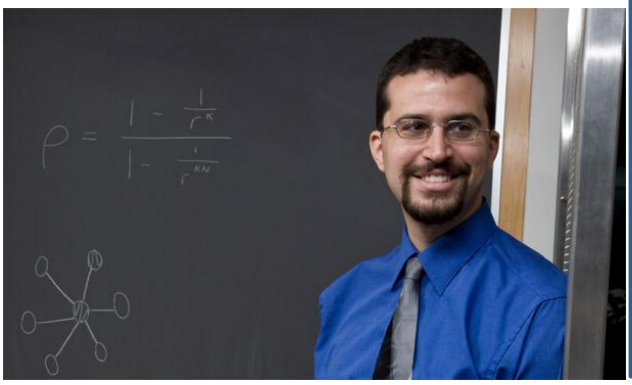
A diferencia de sus predecesores muchos emporios comerciales de hoy no crean registros como meros productos finales de sus líneas de negocio. Google®, Facebook® o Amazon® utilizan herramientas para construir registros digitales personales, históricos. Para estas compañías el registro de la cultura humana es su negocio. Conocen, mejor que el propio cliente, sus preferencias y debilidades. Cuando enviamos un mensaje electrónico nuestros pensamientos o impresiones dejan una huella digital perenne. Google® recordará cada palabra de un agrio c. e. mucho tiempo después de que hayamos olvidado a quién lo enviamos; las fotografías y vídeos en Facebook® serán la crónica de una noche para olvidar. Si escribimos, Google® lo escaneará; si fotografiamos, Flicker® lo almacenará, si un vídeo, YouTube® lo conservará.

Como antecedentes los registros FBIS y SWB, y un artículo previo de los autores antes mencionados. Durante setenta años ─escribe K. Leetaru─ el *Foreing Broadcast Information Service* (FIBS) rastreó las ondas de radio y otros medios de noticias de todo el mundo transcribiendo y traduciendo contenido selecto al inglés creando un inmenso archivo histórico de los medios globales de noticias. El servicio proporcionado por la *Bristish Broadcasting Corporation* (BBC), conocido como *Summary of World Broadcast* (SWB) opera de manera similar. Ambas acciones conocidas como *Open Source Intelligence* (OSINT), cuyas características han sido la obtención en tiempo real de la información y el relativo fácil y mínimo riesgo de su adquisición y utilización. Conocido popularmente como America's window on the world, el FBIS ha sido la columna vertebral de la colección OSINT en la Inteligencia americana. Este material no ha sido muy utilizado en el contexto de la comunidad académica, quizá por el acceso restringido a su contenido. A pesar de las críticas recibidas por los servicios de inteligencia lo realmente destacable es la importancia del «tiempo real». Por otro lado, Lieberman & Michel, utilizando el corpus CELEX publicaron, en 2007, un estudio sobre la cuantificación de la dinámica evolutiva del lenguaje.

La cantidad anual de datos per cápita se acerca al terabyte ($x10^{12}$). Como colectivo la humanidad produce cinco zetabytes ($x10^{21}$) año. Estas cantidades son *big data*. La humanidad duplica la cantidad de información cada dos años. En Google®, un equipo liderado por el ingeniero de *software* Jeremy Ginsberg observó que la gente incrementaba la frecuencia de consultas sobre la gripe ante el anuncio de una próxima epidemia. Establecieron un sistema de alarma en tiempo real sobre dicha consulta. Funcionó. Detectaron una epidemia antes que los Centros de Control de Enfermedades (CDC) de Atlanta. La utilización de *big data* permite a los investigadores de hoy hacer experimentos que sus predecesores ni soñaron. Culturómica es uno de esos experimentos que han dado en llamarse *ciencia agnóstica*: el *modus operandi* del análisis de megadatos asume que a partir de una colección suficientemente grande y diversa de datos puede contestarse a las preguntas más relevantes del tema. Culturómica es a las humanidades lo que la secuenciación de genomas es a la biología o el Gran colisionador de hadrones (LHC, *Large Hadron Collider*) a la física moderna.

Erez Lieberman Aiden, educado en el judaísmo jasídico ortodoxo, teniendo el inglés como tercera lengua (su lengua materna es el rumano y el hebreo la segunda), pasó por Princeton, Harvard y el MIT; cursó estudios rabínicos, de historia, poesía haiku, matemáticas, filología y biología molecular. Sobre la base de la geometría fractal publicó un trabajo en *Science* proponiendo un modelo por el que el genoma humano ─una doble fila de nucleótidos de dos metros de longitud─ se compacta en una especie de ovillo de unas pocas micras de diámetro manteniendo su estructura y función intactas. Pero su salto a la fama se debe a la publicación, en 2011 y en esa misma revista ─científica por excelencia─, en colaboración con Jean-Baptiste Michel y otra docena de avispados postdoc, de un artículo en el que aparece, por vez primera, la palabra «*culturomics*».

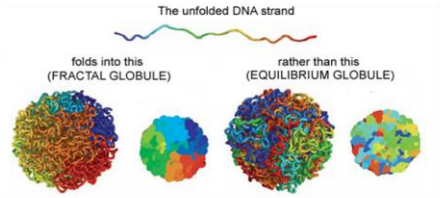
«The renaissance man: how to become a scientist over and over again» (by Ed Young, *Phenomena*).
Erez Lieberman Aiden



«Comprehensive mapping of long-range interactions reveals folding principles of the Human Genome.»

E. Lieberman-Aiden … [18]
Broad Inst. Harvard & MIT.. [13]

***Science* vol. 326, núm. 5950, 9 Oct. 2009**



Science 14 January 2011
Vol. 331 no. 6014 pp.176-182
DOI: 10.1126/science.1199644
Published Online December 16 2010
Received for publication 27 October 2010.
Accepted for publication 6 December 2010.

RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel[1,2,3,4,5,*,†], Yuan Kui Shen[2,6,7], Aviva Presser Aiden[2,6,8], Adrian Veres[2,6,9], Matthew K. Gray[10], The Google Books Team[10], Joseph P. Pickett[11], Dale Hoiberg[12], Dan Clancy[10], Peter Norvig[10], Jon Orwant[10], Steven Pinker[5], Martin A. Nowak[1,13,14], Erez Lieberman Aiden[1,2,6,14,15,16,17,*,†]

Author Affiliations: [1]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA. [2]Cultural Observatory, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Sciences, Harvard University, Cambridge, MA 02138, USA. [4]Department of Psychology, Harvard University, Cambridge, MA 02138, USA. [5]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [6]Laboratory-at-Large, Harvard University, Cambridge, MA 02138, USA. [7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [8]Harvard Medical School, Boston, MA, 02115, USA. [9]Harvard College, Cambridge, MA 02138, USA. [10]Google, Mountain View, CA 94043, USA. [11]Houghton Mifflin Harcourt, Boston, MA 02116, USA. [12]Encyclopaedia Britannica, Chicago, IL 60654, USA. [13]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [14]Department of Mathematics, Harvard University, Cambridge, MA 02138, USA. [15]Broad Institute of Harvard and MIT, Harvard University, Cambridge, MA 02138, USA. [16]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. [17]Harvard Society of Fellows, Harvard University, Cambridge, MA 02138, USA.
[†]To whom correspondence should be addressed. E-mail: jb.michel@gmail.com (J.-B.M.); erez@erez.com (E.L.A.) * These authors contributed equally to this work.

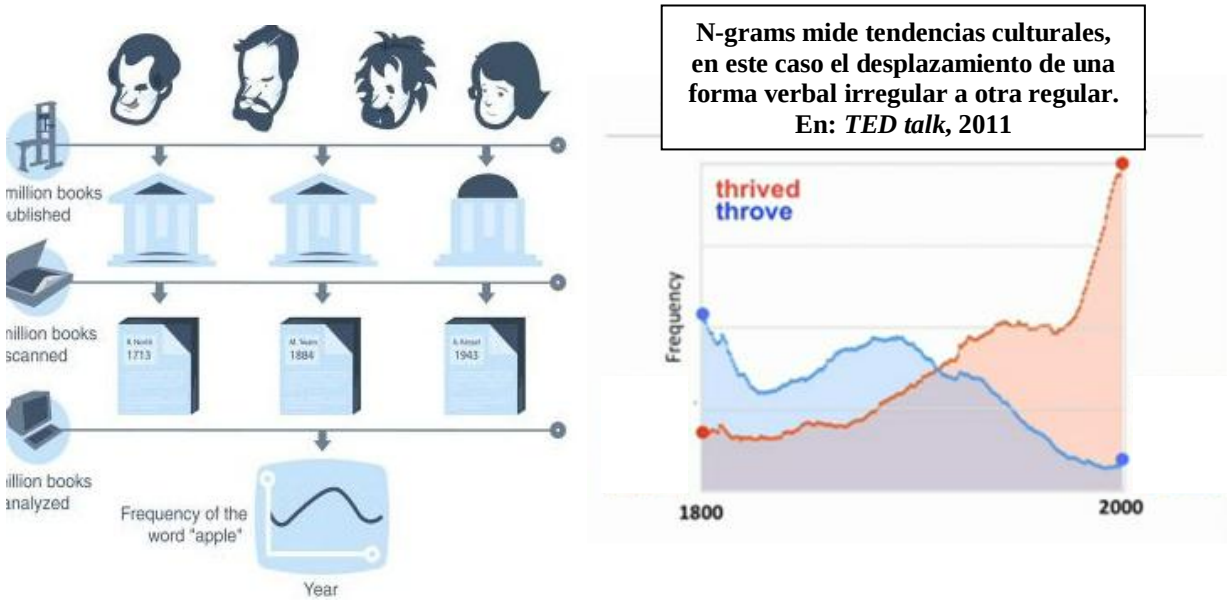
ABSTRACT

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

Cinco millones de libros ─la mayor parte de los libros escaneados [15 millones de libros sobre 130 millones de todos los publicados] son adquiridos directamente por Google de los editores, en especial *Houghton Miffin Harcourt*, Boston, y *Encyclopaedia Britannica*, Chicago; además incorpora otros procedentes de grandes bibliotecas como la de la Universidad de Michigan o la pública de New York─ digitalizados mediante tecnología ROC (reconocimiento óptico de caracteres) y metadatos, conformaron un corpus de más de 500 mil millones (mM) de palabras: 360 mM inglesas, 45 mM francesas, otras 45 mM españolas, 37 mM alemanas, 35 mM rusas, 13 mM chinas y 2 mM hebreas. Los libros más antiguos datan de los años 1500; las primeras décadas están representadas por unas pocas decenas de libros por año que representan unos pocos cientos de miles de palabras. Durante los años correspondientes al periodo de los 1800 el corpus incrementa unos 98 millones de palabras/año; en los 1900, 1.8 mM/año y, a partir de 2000, 11 mM/año.

El corpus resultante fue analizado mediante el innovador *Google n-grams viewer* ®. El estudio inicial comprende los denominados 1-*grams* (de los *n-grams* posibles). 1-*gram* es una secuencia de caracteres entre espacios, lo que incluye palabras ─*house*─ pero también erratas ─*moni*─ y números ─3.1416─. Un *n-gram* es una secuencia de 1-*grams*, tal como las frases «*stock market*» (2-grams) o «*the United States of America*» (5-*grams*). Se denomina 1-*gram* «común» si su frecuencia es superior a uno en mil millones. El estudio inicial restringe los *n-grams* a aquellos 5-*grams* que aparecen más de 40 veces en el corpus. Michel y Aiden se ocupan de la evolución del lexicón inglés, la evolución de la gramática, qué recordamos, lo efímero de la fama o la censura.

Culturómica es la aplicación del análisis de alto rendimiento de megadatos para el estudio de la cultura humana; estrategia que representa un nuevo punto de partida para el análisis histórico más que una sustitución. Mediante el análisis del crecimiento, modificaciones y declinar de las palabras publicadas durante siglos, los matemáticos arguyen que será posible el estudio riguroso de la evolución de la cultura a gran escala utilizando las técnicas matemáticas facilitadas por la biología evolutiva. Otras «disciplinas», como la astrofísica también hacen préstamos a la culturómica; tal es el caso, por ejemplo, de «materia oscura» que en esta última se refiere al léxico común que no se refleja en los diccionarios (el 50%, aproximadamente). Para Anthony Grafton, un historiador de la Universidad de Princeton, se ha iniciado un camino fantástico, aunque añade, «para algunos ─como pasa con frecuencia ante cualquier novedad, puede añadirse─, la aproximación culturómica a las humanidades es algo ajeno, una intromisión». Una poesía puede sentirse, pero también medirse.



N-grams mide tendencias culturales, en este caso el desplazamiento de una forma verbal irregular a otra regular. En: *TED talk*, 2011

En: J-B M *et al. Science* 2011; 331: Fig.1, p. 177.

Niklas Luhmann, en su obra magna *Die Gesellschaft der Gesellscharf*, compara el reto que ha supuesto la emergencia de las computadoras para la sociedad contemporánea con el que supuso, en su momento, la invención de la escritura o de la imprenta; sucesos que revolucionaron las sociedades arcaica y moderna, respectivamente. La idea básica es que una nuevo modo de comunicación permite nuevas formas de autoobservación , lo que al final resulta en una nueva identidad. Sreffen Roth aplica idéntica idea a como el *Google Ngram viewer* ® cambiará conceptos básicos de nuestra sociedad. No es sino un ejemplo del poder de una metodología desarrollada de software que utiliza cantidades ingentes de datos ─*big data* o megadatos─ para resolver problemas mal definidos en ambientes de incertidumbre. En culturómica son datos del lenguaje, pero lo mismo es aplicable a otras situaciones límite. En todos los casos, para evaluar la solución a un problema con parámetros inciertos se necesitan utilizar: conjuntos de entrenamiento ─*training set*─ para construir modelos probabilísticos; conjuntos de validación ─*validation set*─ para desarrollar nuevos algoritmos específicos, y conjuntos de prueba ─*test set*─ para comprobar la robustez de la metodología. Para Peter Norvig, cuantos más datos estén disponibles en la Red y en cuanto la capacidad de computación parece no tener límites, tal metodología orientada hacia datos probabilísticos se convertirá en la principal estrategia para resolver problemas complejos en ambientes de incertidumbre.

El corpus bibliográfico de Google representa un tipo completamente nuevo de megadatos que tiene el potencial de transformar, de manera radical, el estudio del pasado. La mayoría de los repertorios de megadatos son ingentes en cantidad pero cortos en el tiempo; se abastecen de registros recientes de acontecimientos recientes. Ello porque la creación de la fuente subyacente de datos fue catalizada por Internet, una innovación relativamente reciente. Por el contrario, el objetivo de Google no es digitalizar lo contemporáneo sino todo lo escrito desde hace siglos, incluso más atrás de la invención de la imprenta. El corpus googliano ─comentan Aiden y Michel─ no es un simple *big data* sino, además, un *long data* que ha de permitir organizar la información global. Nueve años tras el anuncio del proyecto, en 2004 y ya publicado el artículo seminal, Google había digitalizado más de treinta millones de libros; una colección solo superada por la Biblioteca del Congreso de los EE UU (33 millones). Hoy ya lo habrá superado. Para el 2020 se intuye que se habrá completado la digitalización de todos los libros publicados desde Gutenberg. Faltarán por incluir cartas, manuscritos y todo lo escrito desde la invención de la escritura (ya existe la tecnología). Y también son digitalizables los edificios, las esculturas y cualquier obra de arte.

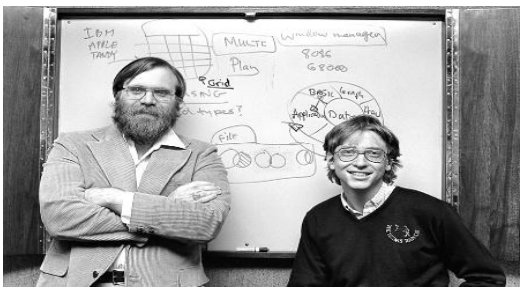
Según Weinberger, los estudios sobre la evolución cultural han sido transformados de manera drástica y en un corto espacio de tiempo, gracias al nivel de accesibilidad, se ha alcanzado un volumen impresionante de datos electrónicos relativos al comportamiento humano. Nuevas estrategias de estudio del uso del lenguaje permiten acceder a una masa de preferencias sociales que inciden de manera directa en la historia en general y en la política en particular que conforman *Culturomics 2.0*. Por ejemplo, el interés por la frecuencia de vocablos y frases afectivas se despertó al observar que la afección social podía predecir las tendencias del mercado; ello se ha seguido de un incremento en el interés del análisis del contenido afectivo de Twitter® como un campo pujante de investigación.

Library of Congress (LOC). 101 Independence Ave., SE. Washington DC 20540

Lieberman Aiden ha dado un paso más allá de los cofundadores de Google ® ─Larry Page y Serguéi Brin─, de Apple ® ─Steve Jobs, Ronald Wayne y Stephen Wozniak─ o de Microsoft ® ─William Henry Gates III (Bill Gates) y Paul Allen─. Todos estos transgredieron las fronteras disciplinares; Lieberman las ignora. Un poema se siente, pero también puede medirse. Se acaban de abrir las puertas a la digitalización de las humanidades. También, el conocimiento innovador, transgresor, ha escapado de la Universidad tradicional. Una reunión académica ─*Shared Horizons*─ en Maryland, en la primavera de 2013, organizada por los *National Institutes of Health*, el *National Endowment for the Humanities* y la *National Library of Medicine*, congregó a un grupo de investigadores interesados desde la historia a las lenguas africanas y a la ciencia de la computación, desde la microbiología a la retórica y la poética a la zoología.

Google®:
Sergey Brin & Larry Page

Apple®:
Steve Jobs, Steve Wozniak & Ronald Wayne



Microsoft®:
Paul Allen & Bill Gates

Lo anterior va de la mano de la capacidad de los individuos para crear, transferir y acceder a la información de manera global. Un grupo internacional de científicos está animado para crear un simulador que pueda replicar todo lo que acontece en el planeta, desde pronósticos meteorológicos y diseminación de brotes epidémicos a transacciones financieras internacionales o contenido de los innumerables SMS (*Short Message Service*). Bautizado como el *Living Earth Simulator* (LES), el proyecto pretende comprender todo lo que sucede en relación con las acciones humanas que perfilan las diferentes sociedades o culturas y el medio ambiente que condiciona el mundo físico. El Dirk Helbing, codirector con Steven Bishop de FuturICT ─megaproyecto que pretende construir el

LES─, comenta que «muchos de los problemas de hoy ─inestabilidades sociales y económicas─ están condicionadas por el comportamiento humano; sin embargo, existe una brecha, hoy por hoy infranqueable, respecto a cómo funcionan la sociedad y la economía». De igual modo que el LHC ha dado un impulso sin precedentes a la física, es necesario un enorme acelerador de conocimiento que provoque la colisión de las diferentes facetas del quehacer humano. Como el LHC o el complejo ALMA (*Atacama Large Millimiter/submillimiter Array*) son ejemplos de la «gran ciencia» a la que se incorporó HUGO (*Human Genome Organization*) ─todos ellos pertenecientes a las «ciencias experimentales»─, LES puede llegar a ser el representante en este «club» de las humanidades y las ciencias sociales. «FuturICT es un proyecto visionario que aportará nueva ciencia y tecnología para explorar, comprender y gestionar nuestro mundo globalmente interconectado. Inspirará nuevas tecnologías de información y comunicación (ICT), socialmente adaptativas e interactivas, que soportaran una inteligencia colectiva». La plataforma FuturICT contempla un sistema nervioso planetario, un simulador Tierra viviente y una plataforma de participación global, que facilitarán una coevolución simbiótica de las TIC y la sociedad. En resumen, un enfoque transcientífico para que seamos capaces de reaccionar a «la ocurrencia de una aceleración acelerada del conocimiento» y contribuir al fortalecimiento de nuestras sociedades. Ello mediante el desarrollo de nuevos enfoques, métodos y tecnologías como el modelado computacional multiescalar, supercomputación social, minería de datos a gran escala o plataformas participativas. «Computación global para nuestro mundo complejo».

Sirva de conclusión la denominada *Wisconsin idea*, atribuida al que fuera Presidente de la Universidad, Charles R. Van Hise, en 1904. Establece que el principio educativo y formativo debe mirar más allá del campus universitario. La *Wisconsin Idea* no es un concepto abstracto. Es, a la vez, una filosofía práctica, una visión a largo plazo, una ruptura epistemológica con el pasado, una actitud y un método. La *Wisconsin Idea* es «*the idealistic and humane concern that knowledge could and should have practical impact on the needs, problems and aspirations of the people*.»

BIBLIOGRAFÍA

Alberto Acerbi, Vasileios Lampos, Philip Garnett, R. Alexander Bantley. «The expression of emotions in 20th century books». *PLoS ONE* **8** (3): e59030. En: http://www.plosone.org/article/ fetchObject.action?uri= info%3Adoi%2F10.1371%2Fjournal.pone.0059030&representation=PDF.

Erez Aiden & Jean-Baptiste Michel. *Uncharted. Big data as a lens on human culture*. Nueva York: Riverhead Books, 2013.

Harold Bloom.. *The Western Canon. The books and schools of the ages.* Nueva York: Harcourt Brace & Co.. Traducción al castellano de Damián Alou -*El Canon Occidental. La escuela y los libros de todas las épocas*- para Editorial Anagrama, Barcelona, 1995.

John Bohannon. «Digital Data. Google opens books to new cultural studies». *Science* 2010 (Dec. 17); **330**: 1600. En: http://dericbownds.net/uploaded_images/Science-2010-Bohannon.pdf.

John Bohannon. «Digital Data. Google books, Wikipedia, and the future of culturomics». *Science* 2011 (Jan. 14); **331** (6014): 135. En: http://www.terceracultura.net/tc/wp-content/uploads/2011/01/culturomics.pdf.

John Bohannon. «The Science Hall of Fame». *Science* 2011 (Jan. 14); **331** (6014): 143. En: http://www.sciencemag.org/content/331/6014/143.3.full.

Johan Bollen, Huina Mao & Xiao-Jun Zeng. «Twitter mood predicts the stock market» *Journal of Computational Science* 2011; **2** (1): 1-8.En: http://arxiv.org/pdf/1010.3003v1.pdf.

Thorsen Brants & Alex Franz. *Web 1T 5-gram Version 1. Linguistic Data Consortium*. University of Pennsylvania, 2006. En: https://catalog.ldc.upenn.edu/LDC2006T13.

Steven Cherry. «The liberal arts goes data mining» *IEEE Spectrum*, 5 Jan 2011. En: http://spectrum.ieee.org/podcast/at-work/education/the-liberal-arts-goes-data-mining.

Steven Cherry. «The cultural treasures in Google Ngram». *IEEE Spectrum*, 9 Jul 2012. En: http://spectrum.ieee.org/podcast/geek-life/profiles/the-cultural-treasures-in-google-ngram

Patricia Cohen. «Analyzing literature by words and numbers». *The New York Times*, Dec. 3, 2010. En: http://www.nytimes.com/2010/12/04/books/04victorian.html?ref=books&pagewanted=print.

Patricia Cohen. «In 500 billion words, new window on culture». *New York Times*, Dec. 16, 2010. En: http://www.nytimes.com/2010/12/17/books/17words.html?_r=0&pagewanted=print.

Culturomics. En: http://www.culturomics.org/. Ver: Susan Haynes «Culturomics: Experimenting on Google books with N-gram viewer», presentación *power-point* en: http://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCEQFjAA&url=http%3A%2F%2Femunix.emich.edu%2F~haynes%2FPapers%2FMASAL2013%2FPresentation1.pptx&ei=gpi3VJTCNoLxUJedgdgI&usg=AFQjCNFcCR0qv_EBKh2v9MTBOXZkssZrTA&sig2=kTq39XHotbke0TvXtHe3VA&bvm=bv.83640239,d.d24

Dead Poets Society. Película dramática norteamericana, estrenada en 1989, producida por Steven Haft, Paul J. Witt & Tony Thomas: escrita por Tom Schulam, dirigida por Peter Weir e interpretada por Robert Williams, para Touchtone Pictures. El ejemplo es utilizado en «Editorial: Poetry in motion»

Liran Einav & Jonathan Levin. «Economics in the age of big data» *Science* 2014; **346** (6210). En: ttp://web.stanford.edu/~leinav/pubs/Science2014.pdf.

Editorial. «Poetry in motion. A quantitative approach to the humanities enriches research». *Nature* 2011 (23 June); **474**: 420. En: http://www.nature.com/nature/journal/v474/n7352/pdf/474420b.pdf.

Michael Erard. «English as she will be spoke.». *New Scientist* 2008 (29 March); 2649: 28-32.

James A. Evans & Jacob G. Foster. «Metaknowledge». *Science* 2011; **331** (6018): 721-725. En: http://www.knowledgelab.org/docs/Metaknowledge.pdf

W. Tecumseh Fitch. «Linguistics: An invisible hand». *Nature* 2007 (11 Oct.); **449** (7163): 665-667. En: http:// www.nature.com/nature/journal/v449/n7163/full/449665a.html.

FuturICT. En: http://www.futurict.eu/sites/default/files/docs/files/FuturICT_32p_Project%20Outline%20WITH%20LHS.pdf.

Denise Gellene. «Evolution of the language: When verbs meet Darwin». *Los Angeles Times* 2007 (Oct. 11). En: http://www.sfgate.com/news/article/Evolution-of-the-language-When-verbs-meet-Darwin -2535667.php.

Anita Gerrini. «Analyzing culture with Google books: Is it social science?» *Pacific Standard. The Science of Society* 2011 (August 7). En: http://www.psmag.com/media/culturomics-an-idea-whose-time-has-come-34742/.

Anthony Grafton. En John Bohannon, 2011.

Eric Hand. «Culturomics: Word play». *Nature* 2011 (23 June); **474** (7352): 436-440. En: http://www.nature.com/news/2011/110617/pdf/474436a.pdf.

Brooks Hanson, Andrew Sugden, Bruce Albers «Editorial: Making data maximally available»; Science staff «Special section: Dealing with data» *Science* 2011, 331 (6018): 649, 692-729. En: http://www.sciencemag.org/content/331/6018.toc

Harvard University Press. «Culturomics, close reading, and casaubon». *HUP Blog* 2011 (June 29). En: http://harvardpress.typepad.com/hup_publicity/2011/06/culturomics-close-reading-and-casaubon. html.

Brian Hayes. «Computer Science - Bit Lit. With digitized text from five million books, one is never at a loss for words» *American Scientist* 2011; **99** (3): 190. En: http://www.americanscientist.org/issues/issue.aspx?id=12418&y=2011&no=3&content=true&page=8&css=print.

Brian Hayes. «With digitized text from five million books, one is never at a loss for words» *American Scientist* 2011; **99** (3): 190. En: http://www.americanscientist.org/issues/num2/bit-lit/1.

James M. Hughes, Nicholas J. Foti, David C. Krakauer & Daniel N. Rockmore. «Quantitative patterns of stylistic influence in the evolution of literature». *PNAS* 2012; **109** (20): 7682-7686. En: http://www.pnas.org/content/109/20/7682.full.pdf+html.

Rudi Keller. *On language change. The invisible hand in language*. (Traducido por Brigitte Nerlich). Londres & N.Y.: Routldge, 1994. En: http://copyfight.me/Acervo/livros/KELLER,%20Rudi.%20On%20Language%20Change%20-%20The%20Invisible%20Hand%20in%20Language.pdf.

Kalev H. Leetaru. «Culturomics 2.0: Forecasting large-scale human behaviour using global news media tone in time and space». *First Monday* 2011 (Sept. 5); **16** (9). En: http://firstmonday.org/ojs/ index.php/fm/rt/printerFriendly/3663/3040.

Kalev H Leetaru. «The scope of FBIS and BBC open-source media coverage, 1979-2008. Studies in Intelligence 2010; 54 (1): 17-37. En: »https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/volume-54-number-1/PDFs-Vol.-54-No.1/U-%20Studies%2054no1-FBIS-BBC-Coverage-Web.pdf.

David W. Letcher. «Culturomics: A new way to see temporal changes in the prevalence of words and phrases». *American Institute of Higher Education. The 6th International Conference Proceedings* 2011; 4(1): 228-235. En: http://www.academia.edu/3051573/THE_FIVE_VALUES_OF_PEOPLE_AND_TECHNOLOGY_DEVELOPMENT_INTRODUCING_THE_VALUE_CREATION_MODEL_FOR_ORGANIZATIONAL_COMPETITIVE_ADVANTAGE_FRAMEWORK.

Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. «Quantifyying the evolutionary dynamics of language». *Nature* 2007; **449** (7163): 713-716. En: http://www.nature.com/nature/journal/v449/n7163/full/nature06137.html.

Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander & Job Dekker. «Comprehensive mapping of long-range interactions reveals folding principles of the human genome». *Science* 2009 (Oct. 9); **326** (5950): 289-293. En: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2858594/.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden,Jon Orwant, Will Brockman & Slav Petrov. «Syntactic Annotations for the Google Books Ngram Corpus». *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pag. 169–174, Jeju, Republic of Korea, 8-14 July 2012. En: http://aclweb.org/anthology/P/P12/P12-3029.pdf.

Niklas Luhmann. *Die Gesellschaft der Gesellschaft*. Frankfurt am Main: Suhrkamp, 1997. Traducción al castellano ─*La Sociedad de la Sociedad*─ de Javier Torres Nafarrate, Darío Rodríguez Mansilla, Marco Omelas Esquinca, Rafael Mesa Iturbide & Areli Montes Suárez para Editorial Herder, México, 2007. En: http://books.google.es/books?id=je1wqYVGD3cC&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

Igor L. Markov. «Review: Limits on fundamental limits to computation». *Nature* 2014; **512** (7513): 147-154. En: http://arxiv.org/pdf/1408.3821v1.pdf.

Emma Marris. «The language barrier». *Nature* 2008; **453**: 446-448. En: http://www.nature.com/news/2008/080521/full/453446a.html

Jean-Baptiste Michel, Y.K Shen, A.P. Aiden, A. Veres, M.K. Gray, The Google Books Team, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak & Erez Lieberman Aiden. «Quantitative analysis of culture using millions of digitized books» *Science* 2011 (14 January); **331**: 176-182. En: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279742/.

Moni Naor. «Verification of a human in the loop or Identification via the Turing Test» 1996 (Sept. 13). En: http://www.wisdom.weizmann.ac.il/~naor/PAPERS/human.pdf.

Peter Norvig. «Natural Language Corpus Data». En: *Beautiful Data. The stories behind elegant data solutions*. Toby Segaran & Jeff Hammerbacher, ed. Beijin: O'Reilly, 1994. Cap. 14; pág. 219-242. En: https://docs.google.com/gview?url=http://it-ebooks.info/read.php?id%3D209-1414841459-7c095def2f 4090558e27c4cf62c 906b6&chrome=true.

Mark Pagel. «Human language as a culturally transmitted replicator». *Nature Reviews Genetics* 2009 (June); **10**: 405-415. En: wiki.helsinki.fi/download/attachments/59060416/Language...

Mark Pagel, Quentin D. Atkinson & Andrew Meade. «Frequency of Word-use predicts rates of lexical evolution throughout Indo-European history». *Nature* 2007 (Oct. 11); **449**: 717-720.

Marco Panza, Domenico Napoletani, Daniele Struppa. «Agnostic Science. Towards a philosophy of data analysis» *Foundations of Science* 2011; **16**: 1-20. En: http://halshs.archives-ouvertes.fr/ docs/ 00/48/32/88/PDF/ Agnostic_Corrected.pdf.

Simon N. Patten, *The New Basis of Civilization*. Daniel M. Fox. ed. John Harvard Library Book, Harvard University Press, 1907.

Matjaž Perc. «The Matthew effect in empirical data». *Journal of The Royal Society Interface* **11**: 20140378. En: http://rsif.royalsocietypublishing.org/content/11/98/20140378.full.pdf.

Alexander M. Petersen, Joel Tenenbaum, Shlomo Havlin & H. Eugene Stanley. «Statistical laws governing fluctuations in word use from word birth to word». *Scientific Reports* 2012; **2**:313 - 1-9. En: http://www.nature.com/srep/2012/120315/srep00313/pdf/srep00313.pdf.

Project Gutenberg Digital Library. En: http://www.gutenberg.org/wiki/Gutenberg:About.

Steffen Roth. Fashionable Functions: A Google Ngram View of Trends in Functional Differentiation (1800-2000). *International Journal of Technology and Human Interaction* 2014; 10 (2): 34-58. En: http://works. bepress. com/cgi/viewcontent.cgi?article=1019&context=roth

Christopher Shea. «The new science of the birth and death of words. Have physicists discovered the evolutionary laws of language in Google's library?» *The Wall Street Journal* 2012 (March 16). En: http://online.wsj.com/news/articles/SB10001424052702304459804577285610212146258#printMode.

Beth A. Simmons.« International Studies in the Global Information Age». *International Studies Quarterly*, 2011; 55(3): 589-599. En: Digital access to scholarship at Harvard: http://dash.harvard. edu/bitstream/handle/1/11365879/simmons-global-information.pdf? sequence=3

Ray Smith. «An Overview of the Tesseract OCR Engi**ne**». 2007 *IEEE*. En: http://tesseract-ocr. google code. com/svn/ trunk/doc/tesseracticdar2007.pdf.

Special report: Managing information. «The Data Deluge». *The Economist* 2010 (Feb. 27).

Julie Steenhuysen. «Irregular verbs: Use them or lose them: study». *Reuters* ed. Chicago, 2007 (Oct. 10). En: http://www.reuters.com/article/2007/10/10/us-language-evolution-idUSN1025847720071010.

TED talk. Jean-Baptiste Michel + Erez Lieberman Aiden: *What we learned from 5 million books*. Filmed Jul 2011. En: http://www.ted.com/talks/what_we_learned_from_5_million_books

David Weinberger. «The machine that would predict the future». *Scientific American* 2011; Dec.: 52-57. En: http://www.cs.virginia.edu/~robins/The_Machine_that_would_Predict_the_Future.pdf.

Wisconsin idea. Charles McCarthy: *The Wisconsin Idea*. New York: Macmillan, 1912. En: http://digicoll. library.wisc.edu/WIReader/Contents/Idea.html

Chris Woodford. «Optical character recognition (OCR)». *Remark Office OMR* 2014 (Oct. 2). En: http://www. explainthatstuff.com/how-ocr-works.html.

Ed Young. «The renaissance man: how to become a scientist over and over again». *PHENOMENA* a science salon hosted by National Geographic Magazine; June 8, 2001. En: http://phenomena. nationalgeographic.com/2011/06/08/the-renaissance-man-how-to-become-a-scientist-over-and-over-again/

Ben Zimmer. «When physicists do linguistics. Is English ‘cooling’? A scientific paper gets the cold shoulder». *The Boston Globe* 2013 (Feb. 10). En: http://www.bostonglobe.com/ideas/2013/02/10/when-physicists-linguistics/ZoHNxhE6uunmM7976nWsRP/story.html.