Research Trends

Special Issue on Big Data

Welcome to the 30th issue of Research Trends.

Research Trends is proud to present this Special Issue on the topic of Big Data. Big Data refers to various forms of large information sets that require special computational platforms in order to be analyzed. This issue looks at the topic of Big Data from different perspectives: grants, funding and science policy; data and computational infrastructure; arts and humanities, and bibliometrics.

Prominent researchers from different institutions and disciplines have been invited to write about the use of Big Data and analytics in their work, providing us with examples of tools, platforms and models of decision making processes. The Special Issue opens with an overview by Gali Halevi and Henk Moed exploring the evolution of Big Data as a scientific topic of investigation in an article that frames the topic within the peer reviewed literature.

Overview of the issue's contributions:

1. Grants, Funding and Science Policy
Julia Lane, a former NSF director and presently senior managing economist at American Institutes for Research, illustrates how Big Datasets such as grants information, authors' networks and co-authorships should be used to inform funding and science policy decisions.

Norman Braveman, an expert in grants writing and the president of BioMed Consultants, demonstrates how sophisticated text mining technologies can be used to analyze big bodies of literature to inform portfolio and gap analysis in an institution's grants applications processes.

Ray Harris, professor at the Department of Geography University College London and Chair of the ICSU Strategic Committee for Information and Data (SCID) writes about the challenges of Big Data and how the International Council for Science (ICSU) sees its approach to Big Data analytics as a way to develop the capability of science to exploit the new era of what is termed "the Fourth Paradigm".

2. Data and computational infrastructure
Daniel Katz and Gabrielle Allen, professors of Computer Science at Louisiana State University, demonstrate how Big Data analytics was enabled at a university level, encouraging the development of infrastructure and strategic approaches to Big Data analytics for universities' depositories.

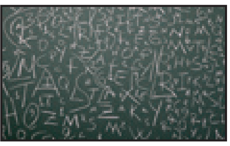
3. Arts and Humanities
Kalev Leetaru, professor at the Graduate School of Library and Information Science at the University of Illinois, shares an innovative way to analyze Wikipedia's view of world history using a Big Data approach to historical research. This article is an extensive cover of the background, methodologies and results of a major project using Wikipedia data. The article is presented here in three sections: (1) background, (2) methodologies and (3) results.

4. Bibliometrics
Research measurements and evaluations using Big Datasets are treated by Henk Moed. In his article he illustrates how usage, citations, full text, indexing and other large bibliographic datasets are being combined and analyzed to follow scientific trends, the impact of research and unique uses of information artifacts in the scientific community.

We hope you enjoy this Special Issue -- please do share your thoughts and feedback with us! You can do this in the comments section following each article on our website www.researchtrends.com or send us an email at: researchtrends@elsevier.com.

Kind regards
Gali Halevi

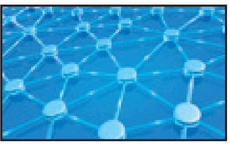
Page 03
The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature
This overview explores the evolution of Big Data as a scientific topic of investigation in an article that frames the topic within the peer reviewed literature.



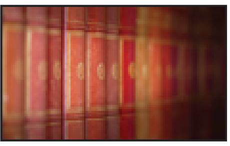
Page 07
Big Data: Science Metrics and the black box of Science Policy
This contribution, by Julia Lane, illustrates how Big Datasets should be used to inform funding and science policy decisions.

Page 09
Guiding Investments in Research: Using Data to Develop Science Funding Programs and Policies
Norman Braveman demonstrates how sophisticated text mining technologies can be used to analyze Big Data.

Page 11
ICSU and the Challenges of Big Data in Science
Ray Harris, discusses challenges of Big Data and ICSU's approach to Big Data analytics.



Page 13
Computational & Data Science, Infrastructure, & Interdisciplinary Research on University Campuses: Experiences and Lessons from the Center for Computation & Technology
Daniel Katz and Gabrielle Allen demonstrate the use of Big Data analytics at university level.



Page 17
A Big Data Approach to the Humanities, Arts, and Social Sciences: Wikipedia's View of the World through Supercomputing
Kalev Leetaru shares an innovative way to analyze Wikipedia's view of world history using a Big Data approach to historical research.

Page 18 – Part 1: Background
This part of the article describes the project background, purpose and some of the challenges of data collection.

Page 21 – Part 2: Data processing and Analytical methodologies
The methods by which the Wikipedia data was stored, processed, and analysed are presented in this part of the article.

Page 24 – Part 3: Data analytics and Visualization
This part of the article describes the analytical methodologies and visualization of knowledge extracted from the Wikipedia data.

Page 31
The use of Big Datasets in bibliometric research
This article illustrates how usage, citations, full text, indexing and other large bibliographic datasets can be combined and analyzed to follow scientific trends.

Page 34
Did you know?
It is our birthday. Happy Birthday **Research Trends**!

Section 1: The Evolution of Big Data as a Research and Scientific Topic

Overview of the Literature.

Gali Halevi, MLS, PhD
Dr. Henk Moed

The term Big Data is used almost anywhere these days; from news articles to professional magazines, from tweets to YouTube videos and blog discussions. The term coined by Roger Magoulas from O'Reilly media in 2005 (1), refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size, but also their complexity. Big Data can be seen in the finance and business where enormous amount of stock exchange, banking, online and onsite purchasing data flows through computerized systems every day and are then captured and stored for inventory monitoring, customer behavior and market behavior. It can also be seen in the life sciences where big sets of data such as genome sequencing, clinical data and patient data are analyzed and used to advance breakthroughs in science in research. Other areas of research where Big Data is of central importance are astronomy, oceanography, and engineering among many others. The leap in computational and storage power enables the collection, storage and analysis of these Big Data sets and companies introducing innovative technological solutions to Big Data analytics are flourishing.

In this article, we explore the term Big Data as it emerged from the peer reviewed literature. As opposed to news items and social media articles, peer reviewed articles offer a glimpse into Big Data as a topic of study and the scientific problems methodologies and solutions that researchers are focusing on in relation to it. The purpose of this article, therefore, is to sketch the emergence of Big Data as a research topic from several points: (1) timeline, (2) geographic output, (3) disciplinary output, (4) types of published papers, and (5) thematic and conceptual development. To accomplish this overview we used Scopus™.

Method

The term Big Data was searched on Scopus using the index and author keywords fields. No variations of the term were used in order to capture only this specific phrase. It should be noted that there are other phrases such as "large datasets" or "big size data" that appear throughout the literature and might refer to the same concept as Big Data. However, the focus of this article was to capture the prevalent Big Data phrase itself and examine the ways in which the research community adapted and embedded it in the mainstream research literature.

The search results were further examined manually in order to determine the complete match between the articles' content and the phrase Big Data. Special attention was given to articles from the 1960s and 1970s which were retrieved using the above fields. After close evaluation of the results set, only 4 older articles were removed from the final results set which left 306 core articles. These core articles were then analyzed using the Scopus analytics tool which enables different aggregated views of the results set based on year, source title, author, affiliation, country, document type and subject area. In addition, a content analysis of the titles and abstracts was performed in order to extract a timeline of themes and concepts within the results set.

Results

The growth of research articles about Big Data from 2008 to the present can be easily explained as the topic gained much attention over the last few years (see Figure 1). It is, however, interesting to take a closer look at older instances where the term was used. For example, the first appearance of term Big Data appears in a 1970 article on atmospheric and oceanic soundings (according to data available in Scopus; see study limitations). The 1970 article discusses the Barbados Oceanographic and Meteorological Experiment (BOMEX) which was conducted in 1969 (2). This was a joint project of seven US departments and agencies with the cooperation of Barbados. A look at the BOMEX site features a photo of a large computer probably used at the time to process the large amounts of data generated by this project (3). Other early occurrences of the term are usually related to computer modeling and software/hardware development for large data sets in areas such as linguistics, geography and engineering.

When segmenting the timeline and examining the subject areas covered in different timeframes, one can see that the early papers (i.e. until 2000) are led by engineering especially in the areas of computer engineering (neural networks, artificial intelligence, computer simulation, data management, mining and storage) but also in areas such as building materials, electric generators, electrical engineering, telecommunication equipment, cellular telephone systems and electronics. From 2000 onwards, the field is led by computer science followed by engineering and mathematics.

Another interesting finding in terms of document types is that conference papers are most frequent followed by articles (see Figures 2 and 3). As we see in the thematic analysis, these conference papers become visible through the abstracts and titles analysis.

The top subject area in this research field is, not surprisingly, computer science; but one can notice other disciplines that investigate the topic such as engineering, mathematics, business and also social and decision sciences (see Figure 4). Other subject areas that are evident in the results sets but not yet showing significant growth are chemistry, energy, arts and humanities and environmental sciences. In the arts and humanities for example, there is a growing interest in the development of infrastructure for e-science for humanities digital ecosystems (for instance, text mining), or in using census data to improve the allocation of funds from public resources.

Finally, we took a look at the geographical distribution of papers. The USA has published the highest number of papers on Big Data by far, followed by China in second place (see Figure 5). In both countries the research on Big Data is concentrated in the areas of computer science and engineering. However, while in the USA these two areas are followed by biochemistry, genetics and molecular biology, in China computer science and engineering are followed by mathematics, material sciences and physics. This observation coincides with other research findings such as the report on International Comparative Performance of the UK Research Base: 2011 (4) which indicated that the USA is strong in research areas such as medical, health and brain research while China is strong in areas such as computer science, engineering and mathematics.

In addition to the overall characteristics of the publications on Big Data, we also conducted a thematic contextual analysis of the titles and abstracts in order to understand how and in what ways the topics within this field have evolved. In order to accomplish this, the abstracts and titles in each article were collected in two batches; one file containing abstracts and titles of articles from 1999-2005 and the second file from 2006-2012. The analysis concentrated on these years rather than the entire set, as there were multiple publications per year during this period. The texts were then entered into the freely available visualization software Many Eyes (www.manyeyes.net).

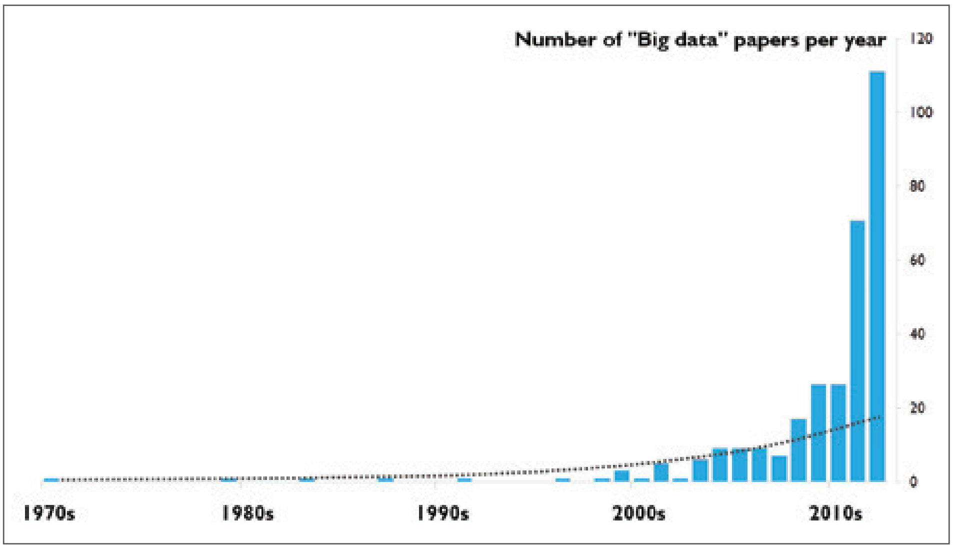


Figure 1: Time line of Big Data as topic of research. The dotted line represents the exponential growth curve best fitting the data represented by the blue bars. This shows the number of Big Data articles increasing faster than the best exponential fit.

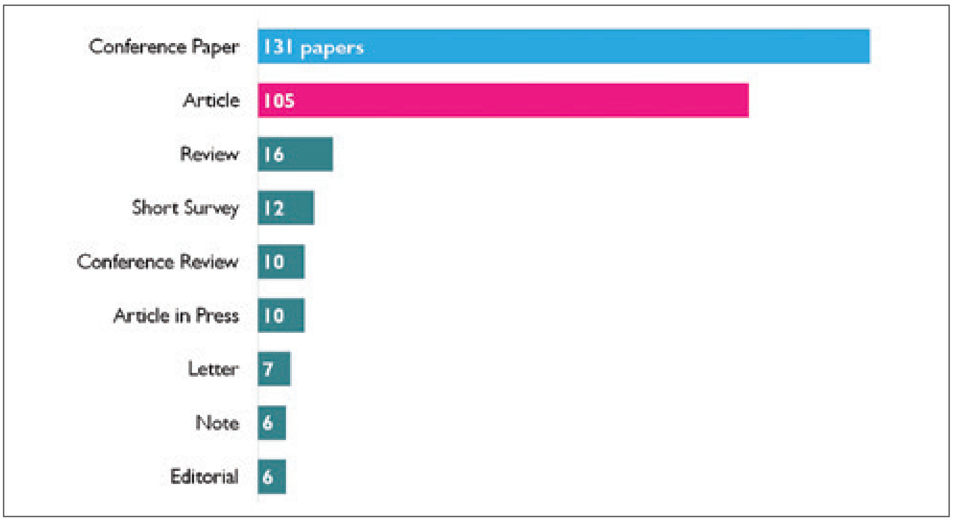


Figure 2: Document types of Big Data papers.

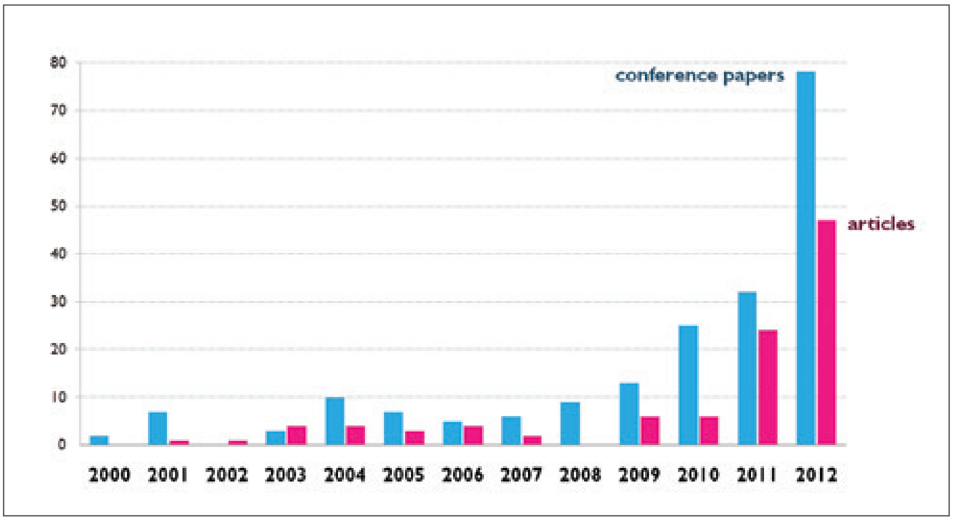


Figure 3: Conference papers and Articles growth over time.

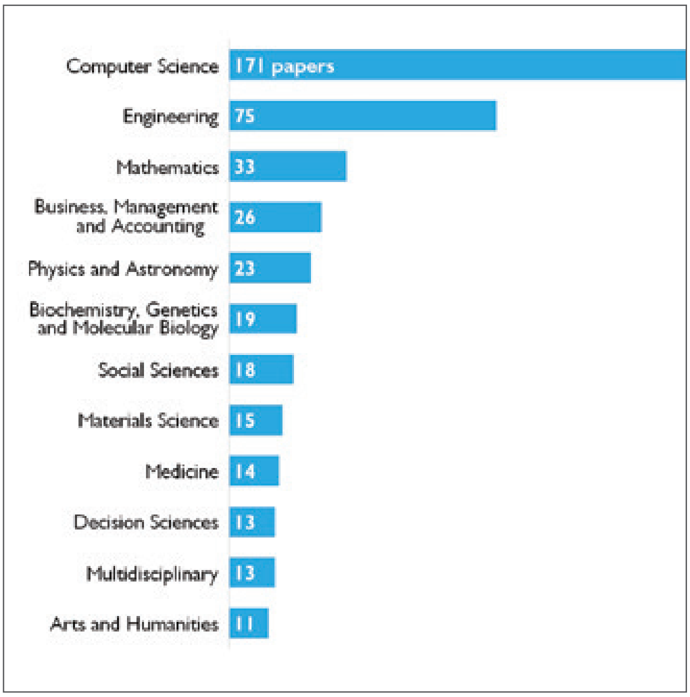


Figure 4: Subject areas researching Big Data.

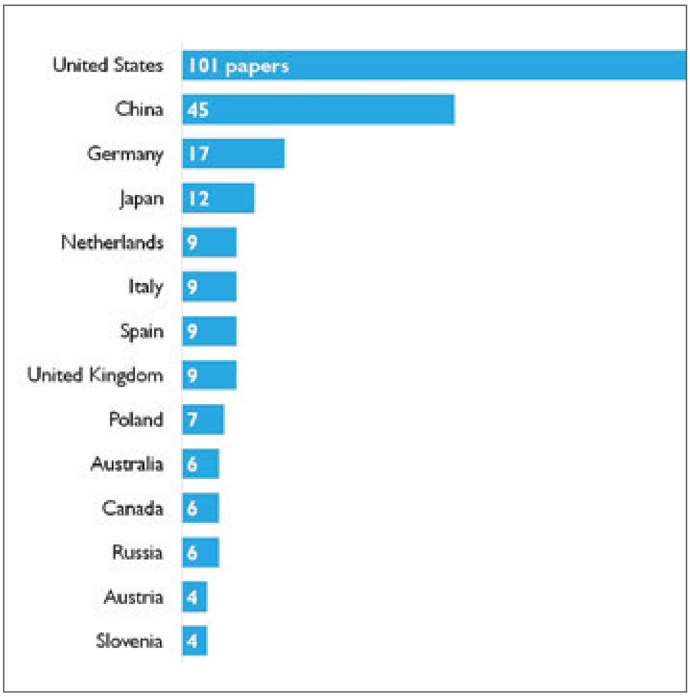


Figure 5: Geographical Distribution of Big Data papers.

This was used to create phrase-maps using the top 50 occurring keywords in these texts. These visualizations were produced by ignoring common words and connecting words such as 'and', 'the', 'of' etc. and used one place space between terms to determine the connections between the terms (see Figures 6 and 7).

These maps visualize two main characteristics of the text: (1) connections between terms are depicted by the gray lines, where a thicker line notes a stronger relationship between the terms; and (2) the centrality of the terms which are depicted by their font size (the bigger the font, the more frequently a term appears in the text). Clusters of connections may appear when a connection is found between single words but not to other clusters.

The first two striking observations when looking at these two maps are the complexity of Figure 6 compared to Figure 7 and the close connectivity of the themes in Figure 7 compared to the scattered nature of their appearance in Figure 6.

The thematic characteristics of the 1999-2005 abstracts and titles text show several scattered connections between two words, seen on the right and left sides of the map. For example, neural networks analysis, on the right side of the map, is a common concept in the field of artificial intelligence computing. This map is conceptually quite simple, with most concepts concentrated around computer related terms, such as 'data mining ',' data sets', XML and applications. When compared to Figure 7 it can be easily seen how the term 'big', although strongly connected to 'data' is not as noticeable as it is in later dates.

The map in Figure 7 represents a tighter network of terms all closely related to one another and to the Big Data concept. Figure 7 also represents a much richer picture of the research on Big Data. There's a clear evolution from basic data mining to specific issues of interest such as data storage and management which lead to cloud and distributed computing. It could be said that the first period demonstrates a naïve picture, in which solutions and topics revolve around a more 'traditional' view of the topic using known concepts of XML and data mining while the second period shows a more complex view of the topic while demonstrating innovative solutions such as cloud computing with emphasis on networks. This also holds for terms such as 'model', 'framework', and 'analytics', that appear in Figure 7, which indicate development and growth in research directions.

A comparison of these two maps also reveals the appearance of diversity in the topics surrounding Big Data such as 'social data' , 'user data' and even specific solutions such as 'MapReduce', a model for processing large datasets implemented by Google (http://mapreduce.meetup.com/), and 'hadoop', an open source software framework that supports data-intensive distributed applications (www.hadoop.apache.org).

As mentioned in the section above analyzing document types, conference papers are central to research in this area. As can be seen in Figure 7, the ACM or IEEE conferences in 2010-2012 play an important role in this area which can be seen by the clear appearance of these terms and their connection to the topic.

Conclusions

Research on Big Data emerged in the 1970s but has seen an explosion of publications since 2008. Although the term is commonly associated with computer science, the data shows that it is applied to many different disciplines including earth, health, engineering, arts and humanities and environmental sciences. Conferences, especially those sponsored by IEEE and/or ACM, are the leaders in the progression of publications in this area followed by journal articles. Geographically, research is led by the USA followed by China and some European countries.

A closer look at the concepts and themes within the abstracts and titles over time show how this area, which began as a computer and technology focus area with some satellite applications, developed into a close and tight-knit discipline featuring applications, methodologies and innovative solutions ranging from could to distributed computing and focusing on user experience. In May 2012, Elsevier sponsored a 2-day conference in Canberra, Australia dedicated to the topics of Big Data, E-science and Science policy (see videos and links to the presentations here: http://www.youtube.com/playlist?list=PL61DD522B24108837). The topic was treated from a variety of viewpoints including the analytics of Big Data sets in publishing, digital scholarship, research assessment and science policy. The multi-dimensional characteristic of this topic is seen in the literature as well as in the social media and online publications. The concept of Big Data as a research topic seems to be growing and it is probable that by the end of 2012 the number of publications will double, if not more, and its analytics and applications will be seen in various disciplines.

Limitations

This study was conducted using Scopus.com in August 2012 and the numbers and percentages presented in this article reflect the indexed publications at the time. These are bound to change as Scopus.com is updated daily with new publications, covering articles in press.

In addition, the dates and document types presented in this study are direct derivatives of Scopus coverage as far as sources and dates. A similar search on other databases might result in slightly different findings and may vary according to the database coverage.

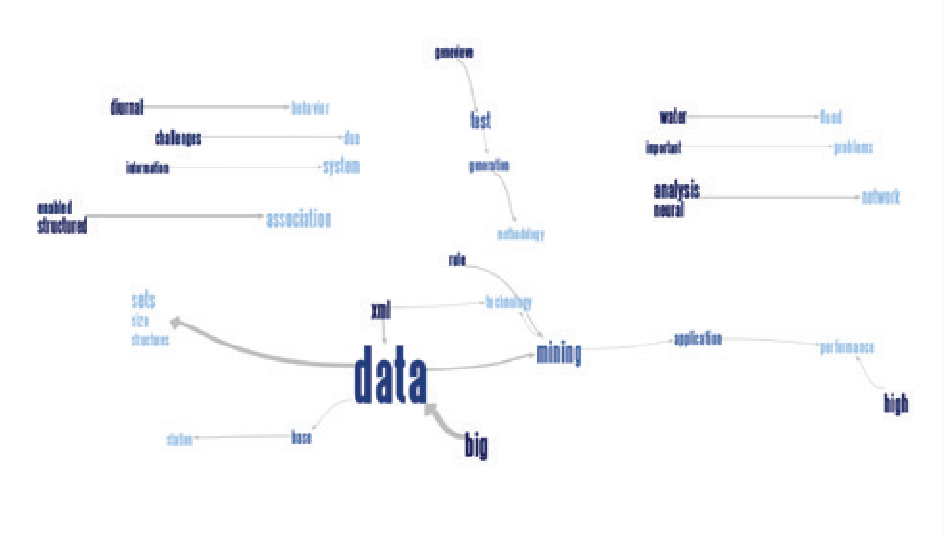


Figure 6: Phrase map of highly occurring keywords 1999-2005.

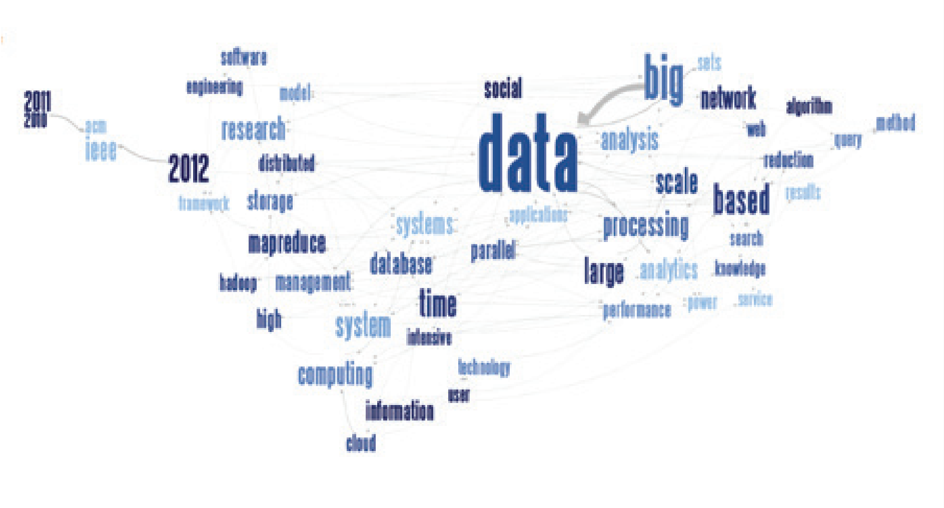


Figure 7: Phrase map of highly occurring keywords 2006-2012.

Useful Links:

1. http://strata.oreilly.com/2010/01/roger-magoulas-on-big-data.html
2. http://www.eol.ucar.edu/projects/bomex/
3. http://www.eol.ucar.edu/projects/bomex/images/DataAcquisitionSystem.jpg
4. http://www.bis.gov.uk/assets/biscore/science/docs/i/11-p123-international-comparative-performance-uk-research-base-2011.pdf

Section 2: **Big Data**

Science Metrics and the black box of Science Policy

Julia Lane
Senior Economics Director,
American Institute of Research

The deluge of data and metrics are generating much heat but shed little light on the black box of science policy. The fundamental problem is conceptual: metrics that connect science funding interventions with numbers of documents miss the key link. Science is done by scientists. Dashboards of metrics that don't link back to scientists are like dashboards missing the link of cables to the engine. They do not provide policy makers with information on how or why funding changed the way in which scientists created and transmitted knowledge. In other words, while bibliometricians have made use of the data deluge to make enormous advances in understanding how to manage scientific documents, the science policy community needs to use the data deluge to make enormous advances in understanding how to manage science (1).

Missing causal links matters

If the focus of funding agencies turns to forcing scientists to produce scientific papers and patents, then they will do so. But if, as the evidence suggests, the essence of science is the creation, transmission and adoption of knowledge via scientific networks, then by missing the causal links, the agencies may distort and retard the very activity they wish to foster. Funding agencies must develop "the ability to define a clear policy intervention, assess its likely impact on the scientific community, find appropriate measures of scientific activities in the pre- and post-period, and define a clear counterfactual." (2) This is no different from Louis Pasteur's swan flask diagram (see Figure 1) that illustrates the fact that spontaneous generation is impossible and that life can never be created out of non-life (3). Like any scientist, we must develop the appropriate conceptual framework that enables us to write down the theory of change of how science policy interventions work – describing what makes the engine run (4).

A sensible organizing framework has been provided by Ian Foster, which identifies individual scientists (or the scientific community consisting of the networks of scientists) as the "engine" that generates scientific ideas. In this case the theory of change is that there is a link between funding and the way in which those networks assemble. Then, in turn, there is a link between scientific networks and the way in which those ideas are created and transmitted, and hence used to generate scientific, social, economic and workforce "products".

Big Data offer science funders a tremendous opportunity to capture those links, precisely because the causal links are often so long and tenuous that relying on manual, individual reporting is, quite simply, destined to fail. The Science of science policy community has been developing a body of knowledge about how to think about and identify those links, rather than just saying,

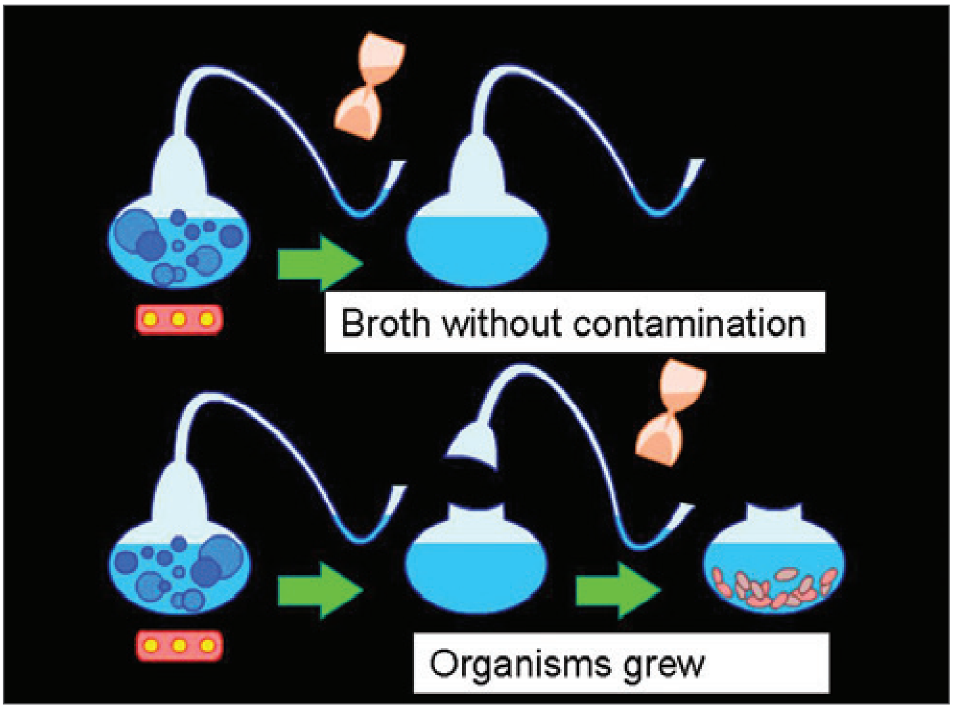


Figure 1: Illustration of swan-necked flask experiment used by Louis Pasteur to test the hypothesis of spontaneous generation.

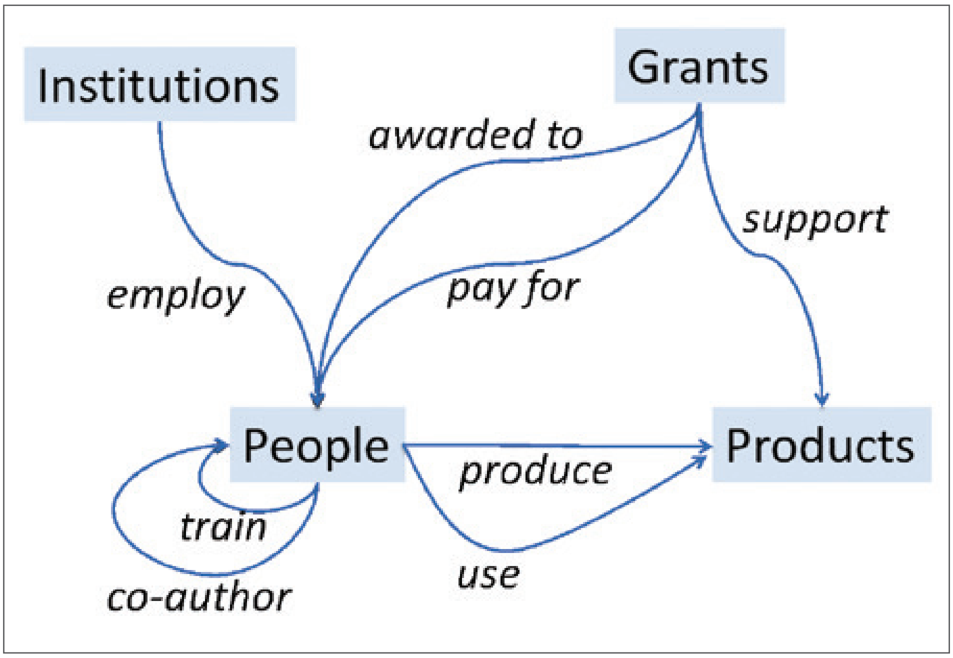
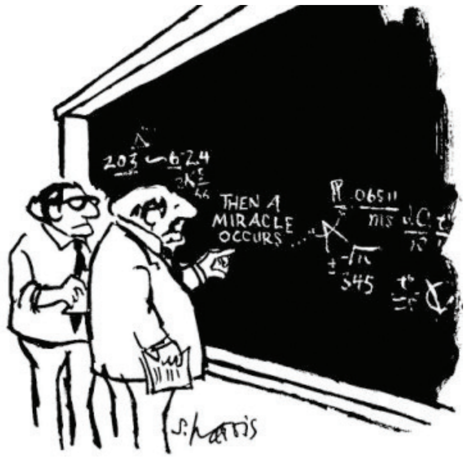


Figure 2: The practice of science (source: Ian Foster).



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Copyright 2006 by Sidney Harris

as the cartoon (left) would have it "that a miracle occurred". The Summer issue of the Journal of Policy Analysis and Management (5), from which the quote was drawn, features articles that document what a science of science policy means in practice; namely, bringing the same intellectual set of models, tools and data to science policy as have been brought to labor, education, and health policy (and many others) (6). The September NSF SciSIP Principal Investigator conference (7) will demonstrate how far this new scientific field has come in moving towards more theoretically grounded metrics – in many cases by both building on the impressive work done by bibliometricians, and working with experts in the field. And the STAR METRICS program has built on the efforts of that community to begin to provide a linked data infrastructure on which those metrics can be founded (8).

In summary, Big Data offers an enormous opportunity to advance the science of science policy. Making the links, so that science funders have new understanding of what is needed to foster science, will enable new light to shine on what has hitherto been a rather black box within which miracles occurred.

References:

1. Lane, J. (2010) "Let's make science metrics more scientific", Nature 464, 488–489.

2. Furman, J. L., Murray, F. & Stern, S. (2012) "Growing Stem Cells: The Impact of Federal Funding Policy on the U.S. Scientific Frontier", J. Pol. Anal. Manage. 31, 661–705.

3. File: Experiment Pasteur English.jpg - Wikipedia, the free encyclopedia. at <http://en.wikipedia.org/wiki/File:Experiment_Pasteur_English.jpg>

4. Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B., & Vermeersch, C.M.J. (World Bank 2011) "Impact Evaluation in Practice", at http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf

5. Journal of Policy Analysis and Management - Volume 31, Issue 3 - Summer 2012 - Wiley Online Library. at <http://onlinelibrary.wiley.com/doi/10.1002/pam.2012.31.issue-3/issuetoc>

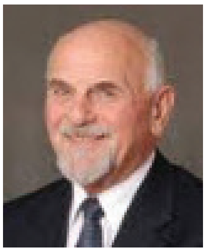
6. Lane, J. & Black, D. (2012) "Overview of the Science of Science Policy Symposium", J. Pol. Anal. Manage. 31, 598–600.

7. NAS CNSTAT SciSIP Principal Investigator Conference. (2012).at <http://www7.nationalacademies.org/cnstat/SciSIP%20Invitation.pdf>

8. Largent, M. A. & Lane, J. I. (2012) "Star Metrics and the Science of Science Policy", Review of Policy Research 29, 431–438.

Section 3: Guiding Investments in Research

Using Data To Develop Science Funding Programs and Policies*



Norman S. Braveman, Ph.D.
President, Braveman BioMed Consultants LLC

One important goal of organizations that provide funds for biomedical and behavioral research is to encourage and support research that leads to more effective health promotion, better disease prevention, and improved treatment of disease. They do this in order to provide a scientific evidence base. In order to ensure that an organization or its programs are effectively moving science forward toward this objective, organizations that fund research must continually assess and re-assess their goals, directions and progress. While there are a variety of ways that funding organizations can carry out these program assessments, there are several discrete and interlinked components common to all approaches including development of: a strategic plan that identifies organizational values, mission, priorities and objectives; an implementation plan listing the timelines, benchmarks, mechanisms of implementation, and the sequence of events related to the elements of the strategic plan; a logic model, based on information gained from all stakeholders, which identifies inputs or available resources that can be used along with expected outcomes from the organization's activities; a gap analysis, an assessment of progress in reaching organizational goals as well as in carrying out the implementation plan by addressing questions about the current position of the organization in relation to where it expected or wanted to be.

In the process of conducting a gap analysis the organization also addresses specific questions about the current state-of-the-science along with pathways to scientific advancement in terms of what is needed to move science ahead, along with identifying barriers to and opportunities for progress. Nevertheless, most program assessments by funding organizations use what I call 'demographic information', that is information that answers questions on the number of grants in a portfolio, how much is being spent on a particular funding program, the mix of grant mechanisms (e.g. basic vs. translational vs. clinical research; investigator initiated vs. solicited research; single project grants vs. large center multi-center grants), and the number of inventions or patents resulting from research supported by any individual portfolio of grants or group of grant portfolios.

While these kinds of measures may be excellent indicators of progress of an organization, with the exception of information about inventions and patents, they seem at least one step removed from measuring the impact of an organization's grant portfolios on the content of science itself. In order to maximize the impact of organizational activities and programs on progress in science, the analysis should use science itself as the data that guides the planning, development and implementation of programs and policies. It's what the scientists whose research the organization may be supporting do in justifying the next step in their research.

There are times when organizations analyze the science of grants in their portfolios by capturing key words in the titles, abstracts, progress reports, and/or grants or grant applications. These are generally tracked over time by program analysts. While the program analysts are typically highly experienced and/or trained, they carry out the analysis by hand and from time to time, from document to document, or from person to person the algorithm they use in classification and categorization can shift in small ways. Such shifts introduce a source of variability that can reduce the reliability and perhaps even the validity of the final results. Moreover, analyzing science by hand is a long, tedious, and expensive task. So our tendency is to do this kind of detailed analysis infrequently…clearly not in 'real time' as seems to be what is needed in this age of fast-paced discovery.

Scientific Fingerprinting

Fortunately, the technology now exists that will allow us to analyze the content of science in a valid, reliable and timely way that overcomes many of the problems that crop up when we do it by hand. More than that, because this approach is computer-based, and therefore fast and reliable, its use allows us to carry out assessments often and on a regular basis. The approach I'm referring to involves the formal textual analysis of scientific concepts and knowledge contained within documents such as strategic and implementation plans, grant applications, progress reports, and the scientific literature. We refer to the output of the textual analysis of a document as a 'scientific fingerprint' or simply a 'fingerprint'.

Without going into the details of the underlying processing, a fingerprint is a precise abstract representation of a text that allows us to look into the text or content rather than only looking at the metadata or demographics. Because fingerprinting is concept driven and not keyword driven and because it uses an ontology (i.e. A is an example of B) as its base, it is not necessary to have a term appear in a document in order for it to be part of a fingerprint. For example, it is possible for a document to contain all of the diagnostic characteristics of a disease but not the name of the disease in order for the disease name to appear in the fingerprint. The only requirement is that the diagnostic characteristics be identified as examples of the named disease somewhere in the scientific literature that makes up the database that is searched. Further, the concepts or weights given to individual concepts comprising a scientific fingerprint of textual content can be adjusted to fit the views of experts in the field. Thus they are not adopted blindly or without validation by experts.

Fingerprinting uses as its base for comparison and analysis the entirety of the Elsevier Scopus database consisting of 45.5 million records, and the information used to develop a fingerprint can be captured relatively easily. Because it is computer-based the textual analysis of the grant portfolio is also much faster than when it is done by hand, thus allowing us to continually assess and reassess an organization's scientific grant portfolio. It is applicable to any document and can be used at any stage of evaluation and program development. In short, it is possible to carry out continual ongoing real time assessments using fingerprinting. Finally, as science changes, as reflected in the scientific literature itself, the fingerprint profile of a given area of science will change, and those changes can be documented and used in an analysis of progress both in science and in the organization's grant portfolio.

While the use of this approach to textual analysis is in its infancy, fingerprinting can allow organizational decision making to be based on the state-of-science, help align organizational goals with the current state-of-the-science, and clarify the organization's role and contributions within a specific area of science. As such, when coupled with demographic data charting the organization's performance it can provide a fuller picture of the current role of the organization in moving science forward, and in the possible role that the organization can play in future scientific development.

*A full version of this paper can be found on Braveman BioMed Consultants'website (http://www.bbmcllc.net/forwhatitsworth.html).

Section 4:
ICSU and the Challenges of Big Data in Science

Professor Ray Harris
Department of Geography
University College London

The Fourth Paradigm

The enormous amounts of data now available to science and to society at large have stimulated some authors to say that we are in the Fourth Paradigm of data-intensive science (1). The First Paradigm was the period of observation, description and experimentation characterised by early scientists and explorers such as Ptolemy and Ibn Battuta. The Second Paradigm was that of the development of theory to explain the way the world works such as in Maxwell's equations and Newton's theory of gravitation and laws of motion. The Third Paradigm developed the earlier theories to create extensive simulations and models such as those used in weather forecasting and in climatology. The reason for the step change to a new paradigm, the Fourth Paradigm, is that the volume of data available to us is so large, now often termed Big Data, that it is both presenting many new opportunities for analysis as well as requiring new modes of thinking, for example in the International Virtual Observatory Alliance (http://virtualobservatory.org) and in citizen science.

Big Data

One clear example of Big Data is the Square Kilometre Array (SKA) (www.skatelescope.org) planned to be constructed in South Africa and Australia. When the SKA is completed in 2024 it will produce in excess of one exabyte of raw data per day (1 exabyte = 10^{18} bytes), which is more than the entire daily internet traffic at present. The SKA is a 1.5 billion Euro project that will have more than 3000 receiving dishes to produce a combined information collecting area of one square kilometre, and will use enough optical fibre to wrap twice around the Earth. Another example of Big Data is the Large Hadron Collider, at the European Organisation for Nuclear Research (CERN), which has 150 million sensors and is creating 22 petabytes of data in 2012 (1 Petabyte = 10^{15} bytes, see Figure 1). In biomedicine the Human Genome Project is determining the sequences of the three billion chemical base pairs that make up human DNA. In Earth observation there are over 200 satellites in orbit continuously collecting data about the atmosphere and the land, ocean and ice surfaces of planet Earth with pixel sizes ranging from 50 cm to many tens of kilometres.

In a paper in the journal Science in 2011, Hilbert and Lopez (2) estimated that if all the data used in the world today were written to CD-ROMs and the CD-ROMs piled up in a single stack, the stack would stretch all the way from the Earth to the Moon and a quarter of the way back again. A report by the International Data Corporation (3) in 2010 estimated that by the year 2020 there will be 35 Zettabytes (ZB) of digital data created per annum.

International Council for Science

The International Council for Science (ICSU) is the coordinating organisation for science and is taking a leading role in developing further the capability of science to exploit the new era of the Fourth Paradigm. The members of ICSU are the 121 national scientific bodies such as the Australian Academy of Sciences and the Royal Society plus the 31 international science unions such as the International Astronomical Union and the International Union for Crystallography. ICSU has always been committed to the principle of the universality of science and in its vision (4) it sees:

"… a world where science is used for the benefit of all, excellence in science is valued and scientific knowledge is effectively linked to policy making. In such a world, universal and equitable access to high quality scientific data and information is a reality …"

Because of its desire to make a reality of universal and equitable access to data, ICSU established three initiatives to address how ICSU can encourage better management of science data (5).

- Panel Area Assessment on Scientific Information and Data, 2003–2004.
- Strategic Committee on Information and Data, 2007–2008.
- Strategic Coordinating Committee on Information and Data, 2009–2011.

World Data System

One of the main outcomes of these ICSU initiatives is the establishment of the World Data System (WDS). In 1957 during the International Geophysical Year (IGY) several World Data Centres were initiated to act as repositories for data collected during the IGY. The number of these data centres increased over time but they were never fully coordinated. The World Data System is now in the process of rejuvenating these data centres by establishing an active network of centres that practice professional data management. The objectives of the WDS are as follows:

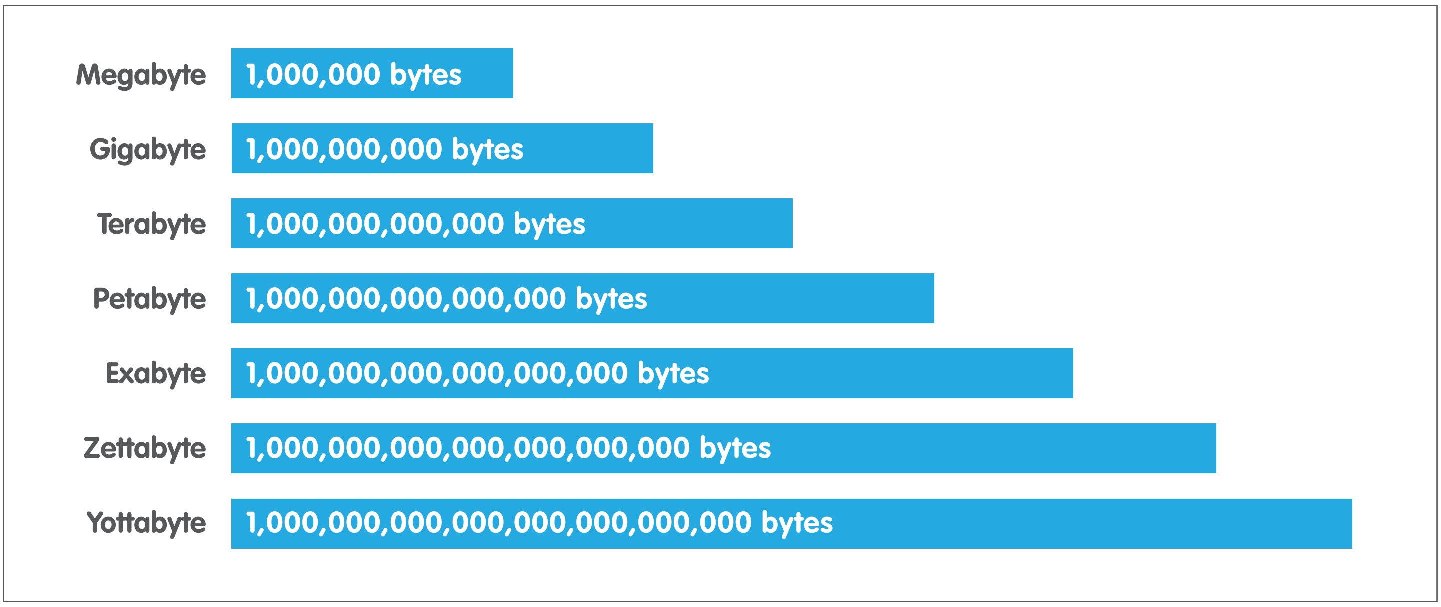


Figure 1: Overview of data scale from megabytes to yottabytes (log scale).

- Enable universal and equitable access to quality-assured scientific data, data services, products and information;
- Ensure long term data stewardship;
- Foster compliance to agreed-upon data standards and conventions;
- Provide mechanisms to facilitate and improve access to data and data products

By early 2012 over 150 expressions of interest in the WDS had been received by ICSU, resulting in over 60 formal applications for membership. Approved members of the World Data System so far include centres for Antarctic data (Hobart), climate data (Hamburg), ocean data (Washington DC), environment data (Beijing) and solid Earth physics data (Moscow) plus the International Laser Ranging Service and the International Very Long Baseline Interferometry Service. By 2013 it is anticipated that the WDS will comprise over 100 centres and networks of active, professional data management.

Further actions

There is still much to do in developing a professional approach to data management in science. The main outstanding issues were addressed by the ICSU Strategic Coordinating Committee on Information and Data noted above and include the following: better guidance for best practice on data management; improved definitions of the various terms used in the phrase "open access"; greater recognition of the publication of data by scientists as well as the publication of journal articles and books; practical help in data management for less economically developed countries through partnership with members of the ICSU family and others; and cooperation with commercial companies for mutual benefit.

Conclusion

Big Data presents science with many challenges, but at the same time presents many opportunities to influence how science grows and develops for the better, not least by adding data-driven science to hypothesis-driven science. Improvements in professional data management will result in better science.

References:

1. Hey, T., Tansley, S. & Tolle, K. (2009) The Fourth Paradigm. "Data-intensive scientific discovery", Microsoft.

2. Hilbert, M. & Lopez, P. (2011) "The world's technological capacity to store, communicate and compute information", Science 332, 1 April 2011, 60-65.

3. IDC (2010) "IDC Digital Universe Study, sponsored by EMC", May 2010, available at http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

4. ICSU Strategic Plan 2006-2011, International Council for Science, Paris, 64pp

5. All the reports are available at the ICSU website, www.icsu.org

Section 5:

Computational & Data Science, Infrastructure, & Interdisciplinary Research on University Campuses:

Experiences and Lessons from the Center for Computation & Technology

Daniel S. Katz [1,2,3] **and Gabrielle Allen** [3,4]

1 Computation Institute,
University of Chicago & Argonne National Laboratory
2 Department of Electrical and Computer Engineering,
Louisiana State University
3 Center for Computation & Technology,
Louisiana State University
4 Department of Computer Science,
Louisiana State University

[This paper is the work of Daniel S. Katz (CCT Director of Cyberinfrastructure Development, 2006 to 2009) and Gabrielle Allen (CCT Assistant Director, 2003 to 2008); it does not reflect the views or opinions of the CCT or LSU.]

Introduction

In recent years, numerous distinguished national panels (1) have critically examined modern developments in research and education and reached a similar conclusion: computational and data-enabled science, as the third pillar of research, standing equally alongside theory and experiment, will radically transform all areas of education, scholarly inquiry, industrial practice, as well as local and world economies. The panels also similarly concluded that to facilitate this transformation, profound changes must be made throughout government, academia, and industry. The remarks made in the 2005 Presidential Information Technology Advisory Committee (PITAC) report (2) are still relevant: "Universities...have not effectively recognized the strategic significance of computational science in either their organizational structures or their research and educational planning." Computational initiatives associated with universities have taken various forms: supercomputing centers that provide national, statewide, or local computing facilities and encourage research involving computation; faculty hiring initiatives focused on initiating research programs to change the university's expertise and culture; establishment of academic research centers on campuses that include formal involvement of faculty, for example, through joint positions with departments; and multi-university or other partnerships where the university is represented by a single entity.

We believe that any academic institution wishing to advance computational and data science needs to first examine its status in three areas: cyberinfrastructure facilities, support for interdisciplinary research, and computational culture and expertise (Figure 1). Cyberinfrastructure facilities refers to the computational, storage, network, and visualization resources (local, national, and international) to which researchers have access; to the technical and professional support for these services; and to the connection of these services to desktop machines or experimental instruments in an end-to-end manner. Support for interdisciplinary research refers to the university's policies on joint appointments between units and associated promotion and tenure, policies and practices for university-wide curricula, and the academic appreciation of computational science that could rate, for example, software or data development in a similar manner to publications and citations. Finally, computational culture and expertise relates to the existence and prominence of faculty across a campus who develop or use computation as part of their research, and the provision of undergraduate and graduate courses that will train and educate students to work on research projects in the computational sciences.

Once the status of these areas has been reviewed, there are additional questions in designing a computational initiative. Should the cyberinfrastructure resources be state-of-the-art to enable leading edge research in computational science? Should faculty expertise in computational science be pervasive across all departments or concentrated in a few departments? Will the university administration back a long-term agenda in computational science and have the sustained desire to implement policies for changing culture? What is the timescale for change?

While there is some literature on issues relating to general interdisciplinary research (e.g. a National Academy review) (3), there is little written on the underlying visions, strategies, issues, practical implementations and best practices for computational initiatives. Further, what exists was usually written for a specific purpose, such as justifying an initiative for a state legislature, funding agency, or campus administration.

Louisiana Experiences

In April 2001, Louisiana Governor Foster asked the state Legislature to fund an Information Technology Initiative as a commitment to the 20-year Vision 2020 plan adopted in 2000 to grow and diversify the state's economy. The legislature authorized a permanent $25 million per year commitment, divided among the state's five research institutions. LSU created the Center for Applied Information Technology and Learning (LSU CAPITAL), targeting funds in education, research, and economic development, with the intent that this investment would result in the creation of new businesses, increased graduates in IT areas, and increased patents and licenses. Edward Seidel was recruited from the Max Planck Institute for Gravitational Physics (AEI) to formulate a vision and detailed plan (4) to structure LSU CAPITAL into a research center related to computation and informational technology, with a physical presence on the campus and a broad mission for interdisciplinary research at LSU and across the state. Seidel became director of LSU CAPITAL, reporting to the LSU vice chancellor of research and economic development. In October 2003, LSU CAPITAL was renamed the LSU Center for Computation & Technology, or CCT (http://www.cct.lsu.edu).

LSU was lacking in all the three areas identified in Figure 1: cyberinfrastructure; support for interdisciplinary research and education; and computational research, which necessitated a three-pronged approach for the center's strategy (5,6).

Cyberinfrastructure

To address LSU's cyberinfrastructure needs, CCT worked to develop campus and regional networks, connect to the national high-speed backbone, and build sustainable computational resources on the campus. (A negative side effect of including a focus on the provision of cyberinfrastructure resources is that some people tend to label the center as just an (Hurricane Protection System) HPC resource provider, rather than a research center; this proved to be an issue with how the center was represented and seen by the LSU administration.) CCT led an effort to propose a statewide high-speed network (called LONI) to connect state research institutions with multiple 10-Gbps optical lambdas. Louisiana Governor Blanco then mentioned LONI as a priority in her State of the State address. At this time, National LambdaRail (NLR) was emerging as a high-speed optical national backbone without a plan to connect to Louisiana. In 2004, Governor Blanco committed $40 million over 10 years to fund LONI, including purchasing and deploying initial computational resources at five sites and supporting technicians and staff, to advance research, education, and industry in the state. The state also funded a membership in NLR to connect the state to computational power available throughout the nation and the world.

When the CCT was formed, LSU had recently deployed what were then significant computational resources: 128-node and 512-node dual-processor clusters, managed by staff from the physics department, and a 46-node IBM Power2/Power3 machine managed by the university's Information Technology Services (ITS). LSU created the HPC@LSU group, funded 50-50 by CCT and ITS to jointly manage these systems, which were the only major compute resources in Louisiana. HPC@LSU also began to manage the LONI compute systems, IBM Power5 clusters, and later, additional Dell systems for both LONI and LSU, including Queen Bee (the largest LONI system), as part of the TeraGrid, the US national HPC infrastructure.

CCT envisioned building a campus and national center for advancing computational sciences across all disciplines, with these groups' research activities integrated as closely as possible with the research computing environment. In this way, the services provided by the computing environment to the campus and nation would be the best possible, and the research output of the faculty, students, and staff would be advanced. CCT faculty would be able to lead nationally visible research activities, being able to carry out a research program that would not be otherwise possible, providing exemplars to the campus, catalyzing activity in computational science approaches to basic sciences, engineering, humanities, business, etc. This was a key component of the CCT vision, one that has been successful at other centers (e.g. NCSA, SDSC, AEI) around the world.

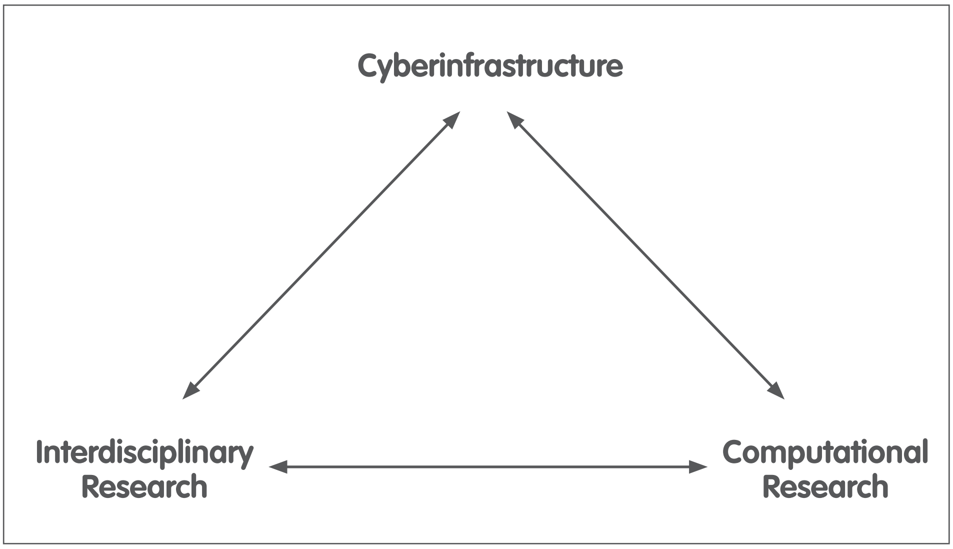


Figure 1: Advancing a comprehensive computational science program requires coordinated initiatives in developing and supporting interdisciplinary research, enabling cyberinfrastructure, and underlying research and culture in computation.

Computational Research

Initially, there were very few computationally oriented faculty in Louisiana, which hindered research in computational science, state collaborations, and LSU's involvement in national or international projects involving computation. To address this, CCT's core strategy has been to recruit computationally-oriented faculty to LSU, generally in joint 50-50 positions with departments, with tenure residing in the departments. This model has been discussed at length and has continuously been seen as the best model for strengthening departments in computational science, and encouraging real buy-in to the overall initiative from the departments. CCT also implements other strategies for associating faculty with the center, both for encouraging and supporting the participation of faculty already on the campus to take an active role in the center's programs and research, and for helping to attract and recruit faculty whose research interests overlap with CCT.

Research staff are also essential, making it possible to quickly bring in expertise in a particular computational area as a catalyst and tool for faculty recruitment, to form a bridge from center activities to the campus, to provide consistent support to strategically important areas, and to facilitate production level software development.

The fundamental group (in the CCT Core Computing Sciences Focus Area), centered around the Departments of Computer Science, Electrical and Computer Engineering, and Mathematics, was to have the necessary skills needed to build and sustain any program in computational science, including computational mathematics, scientific visualization, software toolkits, etc. Application groups were built to leverage strength on campus, hiring possibilities, and new opportunities.

In addition, CCT's Cyberinfrastructure Development (CyD) division aimed to better integrate CCT's research and HPC activities with the campus and national initiatives, with the mission to design, develop, and prototype cyberinfrastructure systems and software for current and future users of LSU's supercomputing systems, partnering where possible with the research groups at CCT to help professionalize prototype systems and support and expand their user base. CyD includes computational scientists, expected to cover 30-50% of their time on proposals led by scientists elsewhere at LSU or LONI, and to spend the rest of their time on computational science activities that lead to new funding or projects and internal support of HPC and LONI activities.

CCT's education goal has been to cultivate the next generation of leaders in Louisiana's knowledge-based economy, creating a highly skilled, diverse workforce. To reach this goal, objectives were set to assist in developing curricula and educational opportunities related to computation, to help hire faculty who would support an integrated effort to incorporate computation into the curricula, to offer programs that support activity in scientific computing, to attract and retain competitive students, and to advance opportunities for women and minorities in the STEM disciplines.

Interdisciplinary Research

The final component of the triangle, interdisciplinary research, was supported by CCT's organization and projects. CCT faculty are generally able to lead and take part in world-class interdisciplinary research groups related to computation, organized in focus areas: Core Computing Sciences, Coast to Cosmos, Material World, Cultural Computing, and System Science & Engineering. Each focus area has a faculty lead responsible for building cross-cutting interdisciplinary research programs, administration, coordinating the hiring of new faculty and staff, and organizing their unit. Interdisciplinary research is driven by activities in strategically motivated, large-scale projects in the focus areas, faculty research groups, and the Cyberinfrastructure Development division. These projects provide support (students, postdocs, and direction) to the Focus Areas as well as broad outreach for education and training across the state. In addition, CCT tried to engage senior administrators and use CCT faculty to drive curriculum change on the campus.

Crosscutting Activities

Two large projects begun in 2007 were the LONI Institute and Cybertools. The LONI Institute was a statewide multi-university collaboration, built on the success of the LONI university partnership, to coordinate the hiring of two faculty members at each university, in computer science, computational biology, and/or computational materials, and of one computational scientist at each university, to spur collaborative projects. Cybertools was another multi-university collaboration that used computational science projects across the state to drive developments in tools that could use the state's computing and networking resources, which in turn could enable new computational science projects.

Particularly from the state legislature's point of view, CCT was intended to catalyze and support new economic development in the state. In fact, the initial metrics for success provided for LSU CAPITAL included the number of resulting new businesses and patents. Economic development needs to be carefully planned and is a long-term initiative, where success can be hard to measure, particularly in the short term. An example success, though not originally planned, was in September 2008, when Electronic Arts (EA) announced that they would place their North American quality assurance and testing center at LSU, creating 20 full-time jobs and 600 half-time jobs, with an annual payroll of $5.7 million throughout the next two years. EA noted that education and research efforts at LSU, including CCT research areas, were a strong factor in the company's decision to locate this center in Louisiana.

Recent Developments and Concluding Thoughts

In 2008, Seidel was recruited to the National Science Foundation, and LSU appointed an interim director and co-director and began a search for a new permanent director, which led to a director being appointed from inside the university for a three-year term. Starting in 2009, LSU has faced several significant and ongoing budget cuts that are currently impacting the CCT, particularly in its ability to recruit and retain faculty and staff.

The issues faced at LSU are similar to those at other institutions responding to the nation's call for an advancement of computation, computational science and interdisciplinary research. We believe it is important to carefully analyze the experiences of centers such as at LSU, as we have attempted to begin to do in this paper, in order to establish best practices for new initiatives or to lead to more fundamental change. From our experiences at CCT, we can highlight four key points that we feel are crucial for the success and sustainability of computational research centers such as CCT:

The three facets of computational science shown in Figure 1 have be taken seriously on the campus at the highest levels and seen as an important component of academic research.

HPC facilities on campuses need to be integrated with national resources and provide a pathway for campus research to easily connect to national and international activities.

Education and training of students and faculty is crucial; vast improvements are needed over the small numbers currently reached through HPC center tutorials; computation and computational thinking need to be part of new curricula across all disciplines.

Funding agencies should put more emphasis on broadening participation in computation, not just focusing on high end systems where decreasing numbers of researchers can join in, but making tools much more easily usable and intuitive and freeing all researchers from the limitations of their personal workstations, and providing access to simple tools for large scale parameter studies, data archiving, visualization and collaboration.

In addition, there are two points that we have learned specifically from the CCT experience:

- The overall vision of the university on topic X needs to be consistent across a broad spectrum of the university administration and faculty; it cannot be just one person's vision, though it may start with one person.

- The funding needs to be stable over a number of years; activities need to be sustained to be successful, and this needs to be clear to the community from the beginning.

References:

1. World Technology Evaluation Center, Inc., (2009) "International Assessment of Research and Development in Simulation-based Engineering and Science", http://www.wtec.org/sbes/

2. President's Information Technology Advisory Committee (2005) "Report to the President of the US, Computational Science: Ensuring America's Competitiveness", http://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf

3. Committee on Facilitating Interdisciplinary Research, National Academy of Sciences, National Academy of Engineering, Institute of Medicine (2004) "Facilitating Interdisciplinary Research", http://www.nap.edu/catalog/11153.html

4. Seidel, E., Allen, G., & Towns, J. "LSU CAPITAL Center (LSUC) Immediate Plans," (2003) http://figshare.com/articles/Original_LSU_CAPITAL_plan/92822

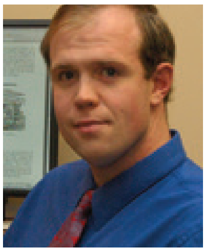
5. CCT Strategic Plan (2006–2010) http://www.cct.lsu.edu/uploads/CCTStrategicPlan20062010.pdf

6. CCT Faculty Plan (2006) http://www.cct.lsu.edu/~gallen/Reports/FacultyPlan_2006.pdf

Section 6:

A Big Data Approach to the Humanities, Arts, and Social Sciences

Wikipedia's View of the World through Supercomputing



Kalev H. Leetaru
Graduate School of Library
and Information Science
University of Illinois

Summary

Wikipedia's view of world history is explored and visualized through spatial, temporal, and emotional data mining using a Big Data approach to historical research. Unlike previous studies which have looked only at Wikipedia's metadata, this study focuses on the complete fulltext of all four million English-language entries to identify every mention of a location and date across every entry, automatically disambiguating and converting each location to an approximate geographic coordinate for mapping and every date to a numeric year. More than 80 million locations and 42 million dates between 1000AD and 2012 are extracted, averaging 19 locations and 11 dates per article and Wikipedia is seen to have four periods of growth over the past millennia: 1001-1500 (Middle Ages), 1501-1729 (Early Modern Period), 1730-2003 (Age of Enlightenment), 2004-2011 (Wikipedia Era).

Since 2007 Wikipedia has hit a limit of around 1.7-1.9 million new mentions of each year, with the majority of its growth coming in the form of enhanced historical coverage, rather than increasing documentation of the present. Two animation sequences visualize Wikipedia's view of the world over the past two centuries, while an interactive Google Earth display allows the browsing of Wikipedia's knowledgebase in time and space. The one-way nature of connections in Wikipedia, the lack of links, and uneven distribution of Infoboxes, all point to the limitations of metadata-based data mining of collections such as Wikipedia and the ability of fulltext analysis and spatial and temporal analysis in particular, to overcome these limitations. Along the way, the underlying challenges and opportunities facing Big Data analysis in the Humanities, Arts, and Social Sciences (HASS) disciplines are explored, including computational approaches, the data acquisition workflow, data storage, metadata construction and translating text into knowledge.

Part 1: Background

This part of the article describes the project background, purpose and some of the challenges of data collection.

Part 2: Data processing and Analytical methodologies

The methods by which the Wikipedia data was stored, processed, and analysed are presented in this part of the article.

Part 3: Data analytics and Visualization

This part of the article describes the analytical methodologies and visualization of knowledge extracted from the Wikipedia data.

Section 6: Part 1

Background

This part of the article describes the project background, purpose and some of the challenges of data collection.

A Big Data exploration of Wikipedia

The introduction of massive digitized and born digital text archives and the emerging algorithms, computational methods, and computing platforms capable of exploring them has revolutionized the Humanities, Arts, and Social Sciences (HASS) disciplines over the past decade. These days, scholars are able to explore historical patterns of human society across billions of book pages dating back more than three centuries or to watch the pulse of contemporary civilization moment by moment through hundreds of millions of microblog posts with a click of a mouse. The scale of these datasets and the methods used to analyze them has led to a new emphasis on interactive exploration, "test[ing] different assumptions, different datasets, and different algorithms … Figure[ing] out whether you're asking the right questions, and … pursuing intriguing possibilities that you'd otherwise have to drop for lack of time."(1) Data scholars leverage off-the-shelf tools and plug-and-play data pipelines to rapidly and iteratively test new ideas and search for patterns to let the data "speak for itself." They are also increasingly becoming cross-trained experts capable of rapid ad-hoc computing, analysis, and synthesis. At Facebook, "on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of [those] analyses to other members of the organization." (1)

The classic image of the solitary scholar spending a professional lifetime examining the most nuanced details of a small collection of works is slowly giving way to the collaborative researcher exploring large-scale patterns across millions or even billions of works. A driving force of this new approach to scholarship is the concept of whole-corpus analysis, in which data mining tools are applied to every work in a collection. This is in contrast to the historical model of a researcher searching for specific works and analyzing only the trends found in that small set of documents. There are two reasons for this shift towards larger-scale analysis: more complex topics being explored and the need for baseline indicators.

Advances in computing power have made it possible to move beyond the simple keyword searches of early research to more complex topics, but this requires more complex search mechanisms. To study topical patterns in how books of the nineteenth century described "The West" using a traditional keyword search, one would have to compile a list of every city and landmark in the Western United States and construct a massive Boolean "OR" statement potentially including several million terms. Geographic terms are often ambiguous ("Washington" can refer both to the state on the West coast and the US capital on the East coast; 40% of US locations share their name with another location elsewhere in the US) and so in addition to being impractical, the resulting queries would have a very high false-positive rate. Instead, algorithms can be applied to identify and disambiguate each geographic location in each document, annotating the text with their approximate locations, allowing native geographic search of the text.

The creation of baselines has also been a strong factor in driving whole-corpus analysis. Search for the raw number of mentions by year of nearly any keyword in a digitized book collection 1800-1900 and the resulting graph will likely show a strong increase in the use of the term over that century. The problem with this measure is that the total number of books published in each year that have been digitized is not constant: it increases at a linear to exponential rate depending on the book collection. This means that nearly any word will show a significant increase in the total number of raw mentions simply because the universe of text has increased. To compensate for this, measurement tools like the Google Ngrams viewer (2) calculate a word's popularity each year not as the absolute number of mentions, but rather as the percentage of all words published that year. This effectively measures the "rate" at which a word is used, essentially normalizing the impact of the increasing number of books each year. Yet, to do this, Google had to compute the total list of all unique words in all books published in each year, creating a whole-corpus baseline. Similarly, when calculating shifts in the "tone" towards a topic or its spatial association, corpus baselines are needed to determine whether the observed changes are specifically associated with that topic, or whether they merely reflect corpus-wide trends over that period.

Into this emerging world of Big Data Humanities, Arts, and Social Sciences (HASS) scholarship, a collaboration with supercomputing company Silicon Graphics International (SGI) leveraged their new 4,000-core 64TB-shared-memory UV2 supercomputer to apply this interactive exploration approach to telling the story of Wikipedia's chronicle of world history. Launched a little over a decade ago, Wikipedia has become an almost indispensable part of daily life, housing 22 million articles across 285 languages that are accessed more than 2.7 billion times a month from the United States alone. Today Alexa ranks it the 6th most popular site on the entire web and it has become one of the largest general web-based reference works in existence (3). It is also unique among encyclopedias in that in addition to being a community product of millions of contributors, Wikipedia actively encourages the downloading of its complete contents for data mining. In fact, it even has a dedicated download site containing the complete contents of the site in XML format ready for computer processing (4). This openness has made it one of the most widely-used data sources for data mining, with Google Scholar returning more than 400,000 articles either studying or referencing Wikipedia.

As an encyclopedia, Wikipedia is essentially a massive historical daybook cataloging global activity through history arrayed by date and location. Yet, most of the literature on Wikipedia thus far has focused on its topical knowledge, examining the linking structure of Wikipedia (which pages link to which other pages and what category tags are applied where) or studied a small number of entries intensively (5). Few studies have explored the historical record captured on Wikipedia's pages. In fact, one of the few previous studies to explore Wikipedia as a historical record visualized just 14,000 events cross-linked from entries that had been manually tagged by human contributors with both date and geographic location information (6). No study has delved into the contents of the pages themselves and looked at every location and every date mentioned across all four million English-language entries and the picture of history they yield from the collective views of the millions of contributors that have built Wikipedia over the past decade.

The Big Data workflow: acquiring the data

The notion of exploring Wikipedia's view of history is a classic Big Data application: an open-ended exploration of "what's interesting" in a large data collection leveraging massive computing resources. While quite small in comparison to the hundreds-of-terabytes datasets that are becoming increasingly common in the Big Data realm of corporations and governments, the underlying question explored in this Wikipedia study is quite similar: finding overarching patterns in a large collection of unstructured text, to learn new things about the world from those patterns, and to do all of this rapidly, interactively, and with minimal human investment.

As their name suggests, all Big Data projects begin with the selection and acquisition of data. In the HASS disciplines the data acquisition process can involve months of searching, license negotiations with data vendors, and elaborate preparations for data transfer. Data collections at these scales are often too large to simply download over the network (some collections can total hundreds of terabytes or even petabytes) and so historically have been shipped on USB drives. While most collections fit onto just one or two drives, the largest collections can require tens, hundreds, or even thousands of high-capacity USB drives or tape cartridges. Some collections are simply too large to move or may involve complex licensing restrictions that prevent them from being copied en-mass. To address this, some data vendors are beginning to offer small local clusters housed at their facilities where researchers can apply for an allocation to run their data mining algorithms on the vendor's own data mining cluster and retrieve just the analytical results, saving all of the data movement concerns.

In some cases it is possible to leverage the high-speed research networks that connect many academic institutions to download smaller collections via the network. Some services require specialized file transfer software that may utilize network ports that are blocked by campus firewalls or may require that the receiving machine install specialized software or obtain security certificates that may be difficult at many institutions. Web-based APIs that allow files to be downloaded via standard authenticated web requests are more flexible and supported on most academic computing resources. Such APIs also allow for nearly unlimited data transfer parallelism as most archives consist of massive numbers of small documents which can be parallelized simply by requesting multiple documents at once. Not all web-based APIs are well-suited for bulk transfers, however. Some APIs only allow documents to be requested a page at a time, requiring 600 individual requests to separately download each page of a single 600 page book. At the very minimum, APIs must allow the retrieval of an entire work at a time as a single file, either as a plain ASCII file with new page characters indicating page boundaries (where applicable) or in XML format. Applications used to manage the downloading workflow must be capable of automatically restarting where they left off, since the downloading process can often take days or even weeks and can frequently be interrupted by network outages and hardware failures. The most flexible APIs allow an application to query the master inventory of all works, selecting only those works matching certain criteria (or a list of all documents), and downloading a machine-friendly CSV or XML output that includes a direct link to download each document. Data mining tools are often developed for use on just one language, so a project might wish to download only English language works, for example.

Many emerging projects perform data mining on the full textual content of each work, and thus require access to the Optical Character Recognition (OCR) output (7). However, handwritten works, works scanned from poor-quality originals (such as heavily-scratched service microform), or works that make use of Fraktur or other specialized fonts, are highly resistant to OCR and thus normally do not yield usable OCR output. Some archives OCR every document and include the output as-is, leading to 10MB files of random garbage characters, while others filter poor-quality OCR through an automated or manual review processes. Those that exclude poor-quality OCR should indicate through a metadata flag or other means that the OCR file has been specifically excluded from this work. Otherwise, it is difficult for automated downloading tools to distinguish between a work where the OCR file has been specifically left out and a technical error that prevented the file from being downloaded (and thus should be requeued to try again). For those documents that include OCR content, archives should include as much metadata as possible on the specific organization scanning the work, the library it was scanned from, the scanning software and imaging system, and the specific OCR software and version used. This information can often be used to incorporate domain knowledge about scanning practices or imaging and OCR pipeline nuances that can be used to optimize or enhance the processing of the resultant text.

Yet, perhaps the greatest challenge in the data acquisition process is policy-based rather than technical. Unlike copyright status, for which there are clear guidelines in determining whether a work has entered the public domain (at least in the United States), there are no national policies or recommendations on what content should be made available for data mining. In some cases archives may have received data from a commercial vendor or other source that may permit browsing, but not computational analysis. In others, funding sources or institutional policy may permit data mining only by researchers at the home institution, or grant them exclusive early access. Some archives permit unrestricted data mining on some content and only "non-consumptive" analysis of other material. Yet, despite this varied landscape of access, few archives have written policies regarding data mining or clear guidelines on what material is available for analysis. Most critically, however, while many archives include a flag for each work indicating whether it has entered public domain, no major archive today has a similar flag to indicate whether a work is available for data mining and under what restrictions. This can cause long delays as archives must evaluate which material can be data mined, in some cases having to create policies and manually review content first. As data mining becomes more commonplace, it is hoped that new national and international guidelines will be formed to help standardize the determination process and that archives will begin to include item-level metadata that indicates the availability of an item for data mining to vastly simplify this process.

In part 2 of this article, the author describes the data processing and analytical methodologies applied to the Wikipedia content.

References:

1. Loukides, M. (2010) "What is Data Science?" http://radar.oreilly.com/2010/06/what-is-data-science.html
2. Google books Ngram Viewer. (online). http://books.google.com/ngrams/
3. Wikipedia. (online). http://en.wikipedia.org/wiki/Wikipedia
4. Wikipedia: Database download. (online). http://en.wikipedia.org/wiki/Wikipedia:Database_download
5. Giles, J. (2005) "Special Report: Internet encyclopaedias go head to head." Nature. http://www.nature.com/nature/journal/v438/n7070/full/438900a.html
6. Lloyd, G. (2011) "A history of the world in 100 seconds." Ragtag.info. http://www.ragtag.info/2011/feb/2/history-world-100-seconds/
7. Leetaru, K. (2011) "Data Mining Methods for the Content Analyst: An Introduction to the Computational Analysis of Informational Content." Routledge.

Section 6: Part 2

Data processing and Analytical methodologies

The methods by which the Wikipedia data was stored, processed, and analysed are presented in this part of the article.

Storing the data for processing

Once the data arrives, it must be processed into a format that can be read by the analysis tools. Many collections are stored in proprietary or discipline-specific formats, requiring preparation and data reformatting stages. One large digital book archive arrives as two million ZIP files containing 750 million individual ASCII files, one for each page of each book in the archive. Few computer file systems can handle that many tiny files, and most analysis software expects to see each book as a single file. Thus, before any analysis can begin, each of these ZIP files must be uncompressed and the individual page files reformatted as a single ASCII or XML file per book. Other common delivery formats include PDF, EPUB, and DjVu, requiring similar preprocessing stages to extract the text layers. While XML is becoming a growing standard for the distribution of text content, the XML standard defines only how a file is structured, leaving individual vendors to decide the specific XML encoding scheme they prefer. Thus, even when an archive is distributed as a single XML file, preprocessing tools will be needed to extract the fields of interest. In the case of Wikipedia, the complete four million entry archive is available as a single XML file for download directly from their website and uses a fairly simple XML schema, making it easy to extract the text of each entry.

As the fields of interest are extracted from the source data, they must be stored in a format amenable to data analysis. In cases where only one or two software packages will be used for the analysis, data can simply be converted into a file format they support. If multiple software packages will be used, it may make more sense to convert the data to an intermediate representation that can easily be converted to and from the other formats on demand. Relational database servers offer a variety of features such as indexes and specialized algorithms designed for datasets too large to fit into memory that enable high-speed efficient searching, browsing, and basic analysis of even very large collections, and many filters are available to convert to and from major file formats. Some servers, like the free edition of MySQL, (1) are highly scalable, yet extremely lightweight and can run on any Linux or Windows server. Alternatively, if it is not possible to run a database server, a simple XML format can be developed that includes only the fields of interest, or specialized formats such as packed data structures that allow rapid randomized retrieval from the file. In the case of the Wikipedia project, a MySQL database was used to store the data, which was then exported to a special packed XML format designed for maximum processing efficiency during the large computation phases.

From words to connections: transforming a text archive into a knowledgebase

Documents are inherently large collections of words, but to a computer each word holds the same meaning and importance as every other word, limiting the types of patterns that can be explored in an archive to simply word frequencies. The creation of higher-order representations capturing specific dimensions of that information, recognizing words indicating space, time, and emotion, allow automated analyses to move closer towards studying patterns in the actual meaning and focus of those documents. The first generation of Big Data analysis focused largely on examining such indicators in isolation, plotting the tone of discussion of a topic over time or mapping locations and making lists of persons mentioned in that coverage. Connections among indicators have largely been ignored, primarily because the incredible richness of human text leads to networks of interconnections that can easily reach hundreds of trillions of links from relatively small collections. Yet historical research tends to revolve around these very connections and the interplay they capture between people, places, and dates and the actions and events that relate them. Thus, the grand challenge questions driving the second generation of Big Data research tend to revolve around weaving together the myriad connections scattered across an archive into a single cohesive network capturing how every piece of information fits together into the global picture. This in turn is driving an increasing focus on connections and the enormous theoretic and computational challenges that accompany them. In the case of Wikipedia, mapping mentions of locations and creating timelines of date mentions and tone in isolation can be enlightening, but the real insight comes from coupling those dimensions, exploring how tone diffuses over space through time.

Thus, once a data archive has been assembled, the first stage of the analytical pipeline usually begins with the construction of new metadata layers over the data. This typically involves using various data mining algorithms to extract key pieces of information, such as names or locations, or to calculate various characteristics of the text, such as readability scores or emotion. The results of these algorithms are then saved as metadata layers to be used for subsequent access and analysis of the text. To explore Wikipedia's view of world history, for example, data mining algorithms were needed to translate its large unstructured text corpus into a structured knowledgebase. Each study uses a different set of data mining algorithms aimed at its specific needs, but location in particular is an emerging class of metadata that is gaining traction as a way of understanding information in a new light. Culturomics 2.0 (2) found that location was the single most prominent organizing dimension in a three-decade archive of more than 100 million print and broadcast news reports translated from vernacular languages across nearly every country in the world, appearing on average 200-300 words. In the case of Wikipedia, previous studies of the linking structure of its pages have found that time and space form the two central dimensions around which the entire site is organized (3). Thus, for the metadata construction stage of the Wikipedia project, a fulltext geocoding algorithm was applied to all of the articles to automatically identify, disambiguate, and convert all textual geographic references to approximate mappable coordinates (4). This resulted in a new XML metadata layer that recorded every mention of a location in the text of each article and the corresponding latitude and longitude for mapping. A similar algorithm was used to identify mentions of dates. For example, a reference to "Georgian authorities" would utilize the surrounding document text to determine whether this referred to the country in Europe or the US state, while a mention of "Cairo" would be disambiguated to see whether it referred to the capital of Egypt or the small town in the state of Illinois in the US. Each location was ultimately resolved to a centroid set of geographic coordinates that could be placed on a map, while each date was resolved to its corresponding year.

Wikipedia provides a facility for article contributors to manually annotate articles with mappable geographic coordinates. In fact, content enriched with various forms of metadata, such as the Text Encoding Initiative (TEI) (5) are becoming more commonplace in many archives. The US Department of State has annotated its historical Foreign Relations of the United States collection with inline TEI tags denoting mentions of person names, dates, and locations (6). However, only selected mentions are annotated, such as pivotal political figures, rather than annotating every person mentioned in each document. This can lead to incomplete or even misleading results when relying on collection-provided metadata. In the case of Wikipedia, the human-provided geographic tags primarily focus on Europe and the Eastern United States, leading to a long history of academic papers that have relied on this metadata to erroneously conclude that Wikipedia is US and European-centric. When switching to the content-based spatial data extracted by the fulltext geocoder, it becomes clear that Wikipedia's coverage is actually quite even across the world, matching population centers (7). As an example of the vast richness obtained by moving from metadata to fulltext, the four million English Wikipedia articles contain 80,674,980 locations and 42,443,169 dates. An average article references 19 locations and 11 dates and there is an average of a location every 44 words and a date every 75 words. As one example, the History section of the entry on the Golden Retriever dog breed (8) lists 21 locations and 18 dates in 605 words, an average of a location every 29 words and a date every 34 words. This reflects the critical role of time and location in situating the narratives of encyclopedias.

Sentiment mining was also used to calculate the "tone" of each article on a 200-point scale from extremely negative to extremely positive. There are thousands of dictionaries available today for calculating everything from positive-negative to anxious-calm and fearful-confident (9). All dictionaries operate on a similar principle: a set of words representing the emotion in question is compiled into a list and the document text is compared against this list to measure the prevalence of those words in the text. A document with words such as "awful", "horrific" and "terrible" is likely to be perceived by a typical reader as more negative than one using words such as "wonderful", "lovely", and "fantastic". Thus, by measuring what percentage of the document's words are found in the positive dictionary, what percent are found in the negative dictionary, and then subtracting the two, a rough estimate of the tonality of the text can be achieved. While quite primitive, such approaches can achieve fairly high accuracy at scale.

Computational resources

All of these dimensions must be brought together into an interconnected network of knowledge. To enable this research, SGI made available one of its UV2 supercomputers with 4,000 processing cores and scalable to 64 terabytes of cache-coherent shared memory. This machine runs a standard Linux operating system across all 4,000 cores, meaning it appears to an end user as essentially a single massive desktop computer and can run any off-the-shelf Linux application unmodified across the entire machine. This is very different from a traditional cluster, which might have 4,000 cores, but spread across hundreds of separate physical computers, each running their own operating system and unable to share memory and other resources. This allowed the project to make use of a rapid prototyping approach to software development to support near-realtime interactive ad-hoc exploration.

All of the metadata extraction, network compilation, workflows, and analysis were done using the PERL (10) programming language and the GraphViz (11) network visualization package. PERL is one of the few programming languages designed from the ground-up for the processing and manipulation of text, especially efficiently extracting information based on complex patterns. One of the greatest benefits of PERL is that it offers many high-level primitives and constructs for working with text patterns and as a scripting language it hides the memory management and other complexities of compiled languages. Often the greatest cost of a research project is the human time it takes to write a new tool or run an analysis, and the ad-hoc exploratory nature of a lot of Big Data analysis means that an analyst is often testing a large number of ideas where the focus is simply on testing what the results look like, not on computational efficiency.

For example, to generate the final network map visualizations, a set of PERL scripts were written to rapidly construct the networks using different parameters to find the best final results in terms of coloration, alpha blending, inclusion thresholds, and other criteria. A script using regular expressions and a hash table was used to extract and store an 800 gigabyte graph entirely in memory, with the program taking less than 10 minutes to write and less than 20 minutes to run. Thus, in less than half an hour, a wide array of parameter adjustments and algorithm tweaks could be tested, focusing on the underlying research questions, not the programming implementation. The shared memory model of the UV2 meant the standard Linux GraphViz package, designed for desktop use, could be used without any modifications to render the final networks, scaling to hundreds of gigabytes of memory as needed. Finally, three terabytes of the machine's memory were carved off to create a RAM disk, which is essentially a filesystem that exists entirely in system memory. While such filesystems are temporary, in that they are lost if the machine is powered down, their read/write performance is limited only by the speed of computer memory and is over 1,000 times faster than even traditional solid state disk. In this project, the use of a RAM disk meant that all 4,000 processor cores could be reading and writing the same set of common files in non-linear order and experience little to no delay, whereas a traditional magnetic disk system would support only a fraction of this storage load.

Part 3 of this article presents the results of the analysis including growth trends and visualization.

References:

1. http://www.mysql.com/

2. Leetaru, K. (2011) "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space", First Monday. 16(9). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040

3. Bellomi, F. & Bonato, R. (2005) "Network Analysis for Wikipedia", Proceedings of Wikimania.

4. Leetaru, K. (forthcoming). "Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards a Geographically Enriched Wikipedia", D-Lib Magazine.

5. http://www.tei-c.org/index.xml

6. http://history.state.gov/historicaldocuments

7. Leetaru, K. (forthcoming). "Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards a Geographically Enriched Wikipedia", D-Lib Magazine.

8. http://en.wikipedia.org/wiki/Golden_Retriever

9. Leetaru, K. (2011). "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space", First Monday. 16(9). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040

10. http://www.perl.org/

11. http://www.graphviz.org/

Section 6: Part 3

Data analytics and Visualization

This part of the article describes the analytical methodologies and visualization of knowledge extracted from the Wikipedia data.

The growth of world knowledge

Putting this all together, what can all of this data say about Wikipedia's view of world history? One of the greatest challenges facing historical research in the digital era is the so-called "copyright gap" in which the majority of available digital documents were published either in the last few decades (born digital) or prior to 1924 (copyright expiration). The vast majority of the twentieth century has gone out of print, yet is still protected by copyright and thus cannot be digitized. Computational approaches can only examine the digital record and as scholarship increasingly relies on digital search and analysis methods, this is creating a critical knowledge gap in which far more is known about the literature of the nineteenth century than of the twentieth. In an illustration of how severe a problem this has become, one recent analysis of books in Amazon.com's warehouses found there were twice as many books from 1850 available as digital reprints as there were from 1950 due to this effect (1). It seems logical that perhaps Wikipedia's contributors might rely on digitized historical resources to edit its entries and thus this same effect might manifest itself in Wikipedia's view of history.

Figure 1 shows the total number of mentions across Wikipedia of dates in each year 1001AD to 2011, visualizing its timeline of world history. The date extraction tool used to identify all date mentions works on any date range, but four-digit year mentions are more accurate since in Wikipedia four-digit numbers that are not dates have commas, reducing the false positive rate. Immediately it becomes clear that the copyright gap seen in other collections has not impacted the knowledge contained in Wikipedia's pages. Instead, there is a steady exponential growth in Wikipedia's coverage through time, matching intuition about the degree of surviving information about each decade. For the purposes of this study, references to decades and centuries were coded as a reference to the year beginning that time period ("the 1500's" is coded as the year 1500), which accounts for the majority of the spikes. One can immediately see major events such as the American Civil War and World Wars I and II. Figure 2 shows the same timeline, but using a log scale on the Y axis. Instead of displaying the raw number of mentions each year, a log scale displays exponential growth, making it easier to spot the large-scale patterns in how a dataset has expanded over time. In this case, the log graph shows that Wikipedia's historical knowledge 1001AD-2011 largely falls into four time periods: 1001-1500, 1501-1729, 1730-2003, 2004-2011. During the first period (roughly corresponding to the Middle Ages) the number of mentions of each year has a slow steady growth rate from around 2,200 mentions about each year to around 2,500 a year. This rapidly accelerates to around 6,500 mentions during the second period (corresponding to the Early Modern Period, starting around the late Renaissance), then increases its growth rate once again in the third period (corresponding to the start of the Age of Enlightenment) through 650,000 mentions of each year in the third period. Finally, the fourth period begins with the rise of Wikipedia itself (the "Wikipedia Era"), with a sudden massive growth rate far in excess of the previous periods.

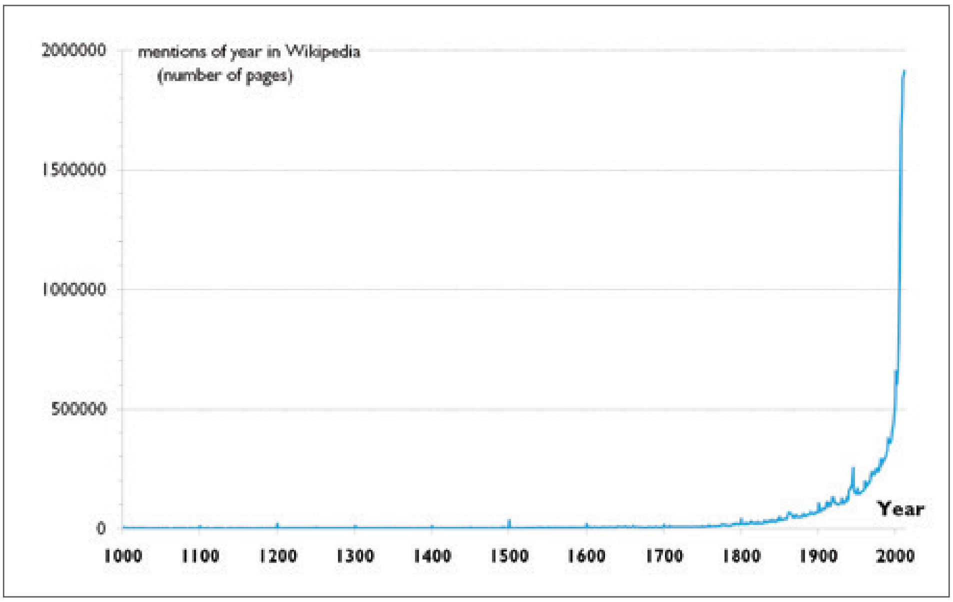


Figure 1: Number of mentions of each year 1001AD-2011 in Wikipedia (Y axis is number of pages).

Figure 3 shows a zoom-in of the period 1950-2011, showing that the initial spike of coverage leading into the Wikipedia Era begins in 2001, the year Wikipedia was first released, followed by three years of fairly level coverage, with the real acceleration beginning in 2004. Equally interesting is the leveling-off that begins in 2008 and that there are nearly equal numbers of mentions of the last three years: 2009, 2010, and 2011. Does this reflect that Wikipedia is stagnating or has it perhaps finally reached a threshold at which all human knowledge generated each year is now recorded on its pages and there is simply nothing more to record? If the latter was true, this would mean that most edits to Wikipedia today focus on contemporary knowledge, adding in events as they happen, turning Wikipedia into a daybook of modern history.

Figure 4 offers an intriguing alternative. It plots the total number of articles in the English-language Wikipedia by year 2001-2011 against the number of mentions of dates from that year. There are nearly as many mentions of 2007 as there were pages in Wikipedia that year (this does not mean every page mentioned that year, since a single page mentioning a year multiple times will account for multiple entries in this graph). Since 2007, Wikipedia has continued to grow substantially each year, while the number of mentions of each of those years has leveled off. This suggests that Wikipedia's growth is coming in the form of enhanced coverage of the past and that it has reached a point where there are only 1.7-1.9 million new mentions of the current year added, suggesting the number of items deemed worthy of inclusion each year has peaked.

Of course, the total number of mentions of each year tells only one part of the story. What was the emotional context of those mentions? Were the events being described discussed in a more negative or a more positive light?

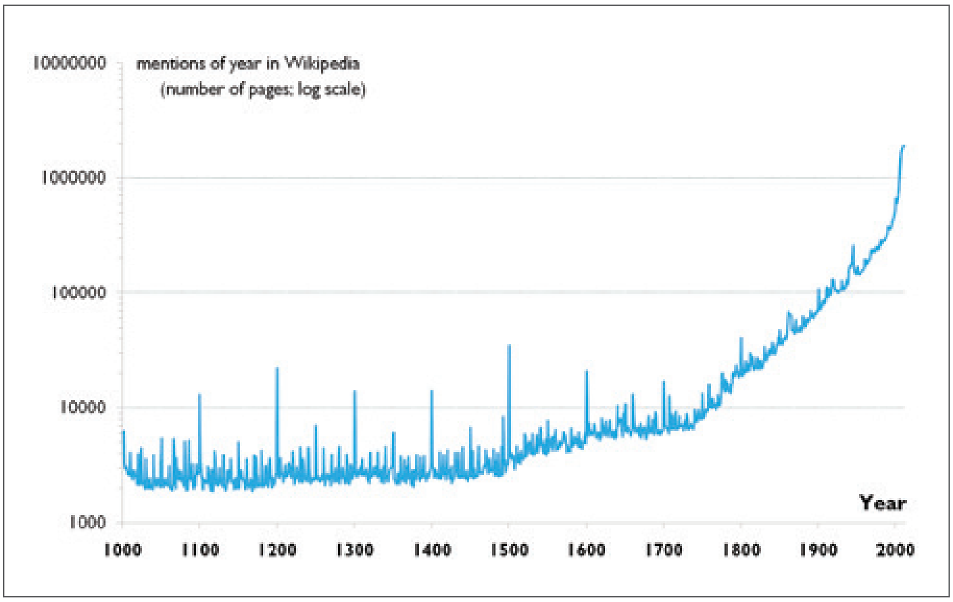


Figure 2: Number of mentions of each year 1001AD-2011 in Wikipedia (Y axis is log scale of page count to show growth rate).

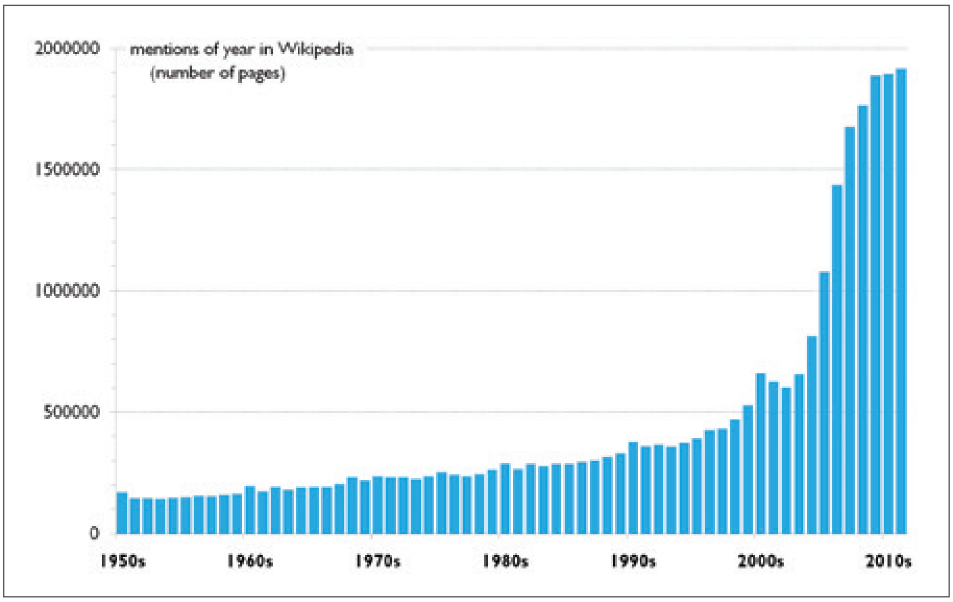


Figure 3: Number of mentions of each year 1950-2011 in Wikipedia (Y axis is number of pages).

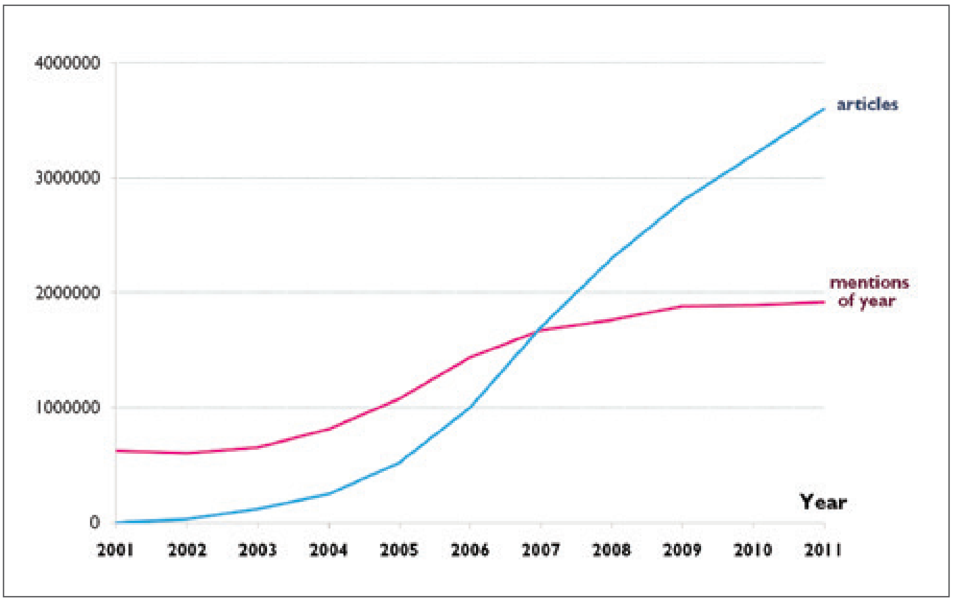


Figure 4: Size of Wikipedia versus number of mentions of that year 2001-2011.

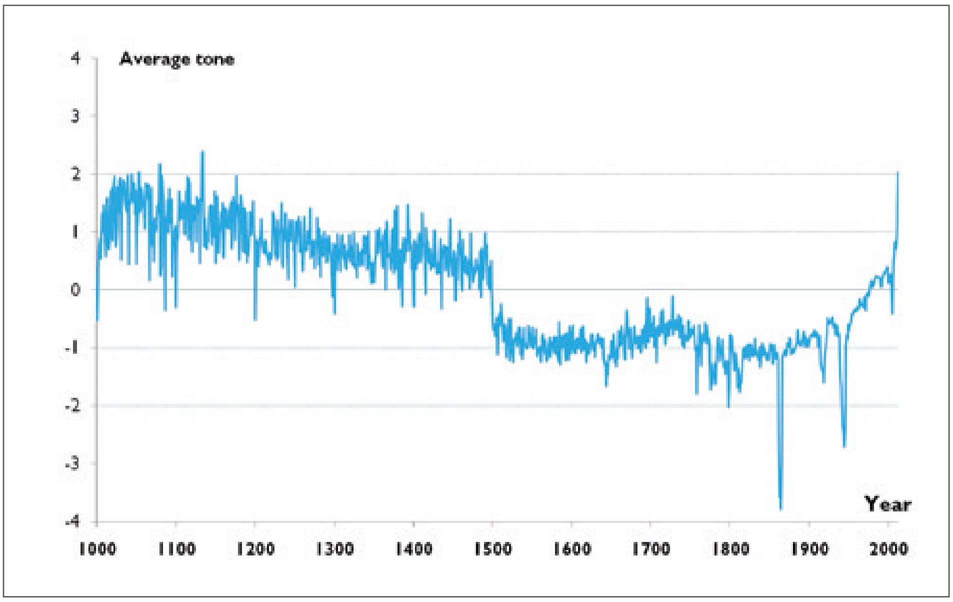


Figure 5: Average tone of all articles mentioning each year 1001AD-2011 (Y axis is Z-score).

Figure 5 visualizes how "positive" or "negative" each year was according to Wikipedia (to normalize the raw tonal scores, the Y axis shows the number of standard deviations from the mean, known as the Z-score). Annual tone is calculated through a very simplistic measure, computing the average tone of every article in Wikipedia and then computing the average tone of all articles mentioning a given year (if a year is mentioned multiple times in an article, the article's tone is counted multiple times towards this average). This is a very coarse measure and doesn't take into account that a year might be referenced in a positive light in an article that is otherwise highly negative. Instead this measure captures the macro-level context of a year: on the scale of Wikipedia, if a year is mentioned primarily in negative articles, that suggests something important about that year.

One of the most striking features of Figure 5 is the dramatic shift towards greater negativity between 1499 and 1500. Tone had been becoming steadily more negative from 1001AD to 1499, shifting an entire standard deviation over this period, but there is a sudden sharp shift of one full standard deviation between those two years, with tone remaining more negative until the most recent half-century. The suddenness of this shift suggests this is likely due to an artifact in Wikipedia or the analysis process, rather than a genuine historical trend such as a reflection of increasing scholarly questioning of worldly norms during that period. Possibilities include a shift in authorship or writing style, or increased historical documentary record that covers a greater class of events. Another striking plunge towards negativity occurs from 1861-1865, reflecting the American Civil War, with similar plunges around World Wars I and II. World War II shows nearly double the negativity that World War I did, but just three quarters of that of the Civil War.

Visualizing Wikipedia over time and space

The Figures above show the power of visualizing Wikipedia temporally, but to really understand it as a global daybook of human activity, it is necessary to add the spatial dimension. The primary geographic databases used for looking up location coordinates are limited to roughly the last 200 years, so here the analysis was limited to 1800-present (2). Each location was associated with the closest date reference in the text and vice-versa, leading to a spatially and temporally-referenced network capturing the locations and connections among those locations through time recorded in Wikipedia's pages. For every pair of locations in an article with the same associated year, a link was recorded between them. The average tone of all articles mentioning both locations with respect to the same year was used to compute the color of that link. A scale from bright green (high positivity) through bright red (high negativity) was used to render tone graphically. The importance of time and location in Wikipedia results in more than 3,851,063 nodes and 23,672,214 connections across all 212 maps from 1800-2012. The massive number of connections meant most years simply became an unintelligible mess of crisscrossing links. To reduce the visual clutter, the first sequence discarded links that appeared in less than 10 articles (see Figure 6). This preserves only the strongest links in the data. To focus only on the linking structure, the second sequence displayed all links, but discarded the tonal information and made each edge semi-transparent so they blended into one another (see Figure 7). The result is that an isolated link with no surrounding links will appear very faint, while lots of links overlapping on top of each other will result in a bright white flare. By focusing purely on the linking structure, this animation shows evolving connections across the world.

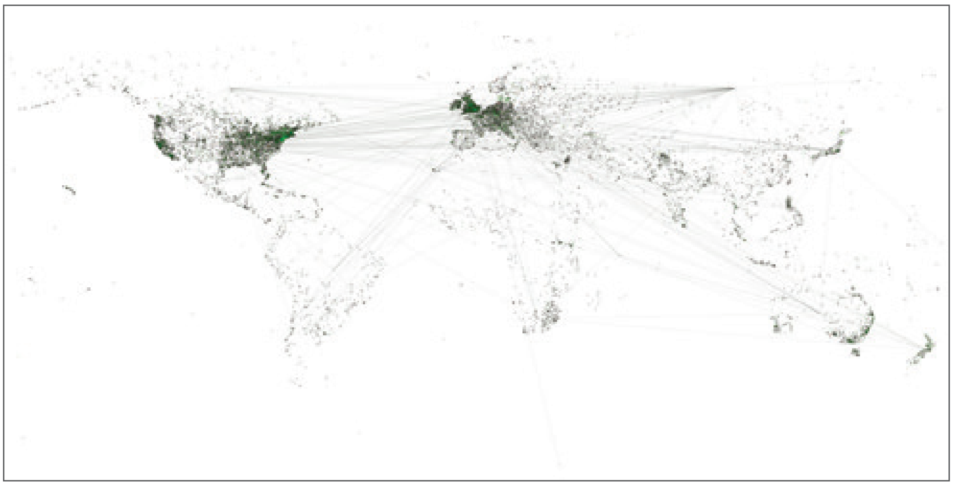


Figure 6: Tone map (see video at https://www.youtube.com/watch?v=KmCQVlVpzWg).

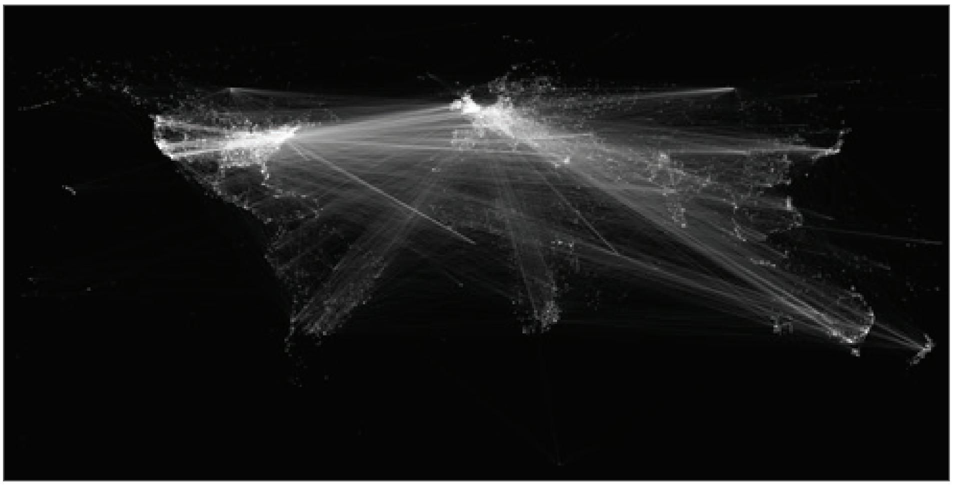


Figure 7: Intensity map (see video at https://www.youtube.com/watch?v=wzuOcP7oml0).

Interactively browsing Wikipedia through time and space

While animations are an extremely powerful tool for visualizing complex information, they do not allow users to interactively drill into the data to explore interesting trends. Ultimately one would like to be able to convert those static images into an interactive interface that would enable browsing Wikipedia through time and space. As an example, let's say one was interested in everything Wikipedia said about a certain area of Southern Libya in the 1840's and 1850's. Wikipedia's own keyword search interface would not be useful here, as it does not support advanced Boolean searches, only searches for a specific entry. Since the Wikipedia search tool does not understand the geographic and date information contained on its pages, one would have to manually compile a list of the name of every city and location in the area of interest, download a copy of Wikipedia, and write a program to run a massive Boolean search along the lines of "(city1name OR city2name OR city3name OR …) AND (1841 OR 1842 OR …)". Obviously such a task would be infeasible for a large area and highly labor-intensive and error-prone even for small queries. This is a fundamental inconsistency of Wikipedia as it exists today: it contains one of the richest open archives of historical knowledge arrayed through time and space, but the only mechanism of interacting with it is through a keyword search box that cannot take any of this information into account.

To prototype what such an interface might look like, all of the information from the animation sequences for Libya 1800 to 2012 described above was extracted and used to create a Google Earth KML file. Figure 8 links to a Google Earth file (3) that offers interactive browsing of Wikipedia's coverage of Libya over this period. Libya was chosen because it offered a large geographic area with a fair amount of change over time, while still having few enough points that could easily load in Google Earth. Unfortunately, most geographic mapping tools today support only a small number of points and Google Earth is one of the few systems that supports date-stamped records. Each location is date-stamped in this demo to the year level so the Google Earth time slider feature can be used to move through time to see what locations of Libya have been mentioned with respect to different time periods over the last 212 years (note that Google Earth operates at the day level, so even though this data is at the year level, Google Earth will show individual days in the time slider).

The display can be narrowed to show only those locations mentioned with respect to a certain timeframe, or one can scroll through the entire 212 years as an animation to see which areas have attracted the attention of Wikipedia's editors over time. Imagine being able to load up the entire world in this fashion and browse all of Wikipedia's coverage in time and space!

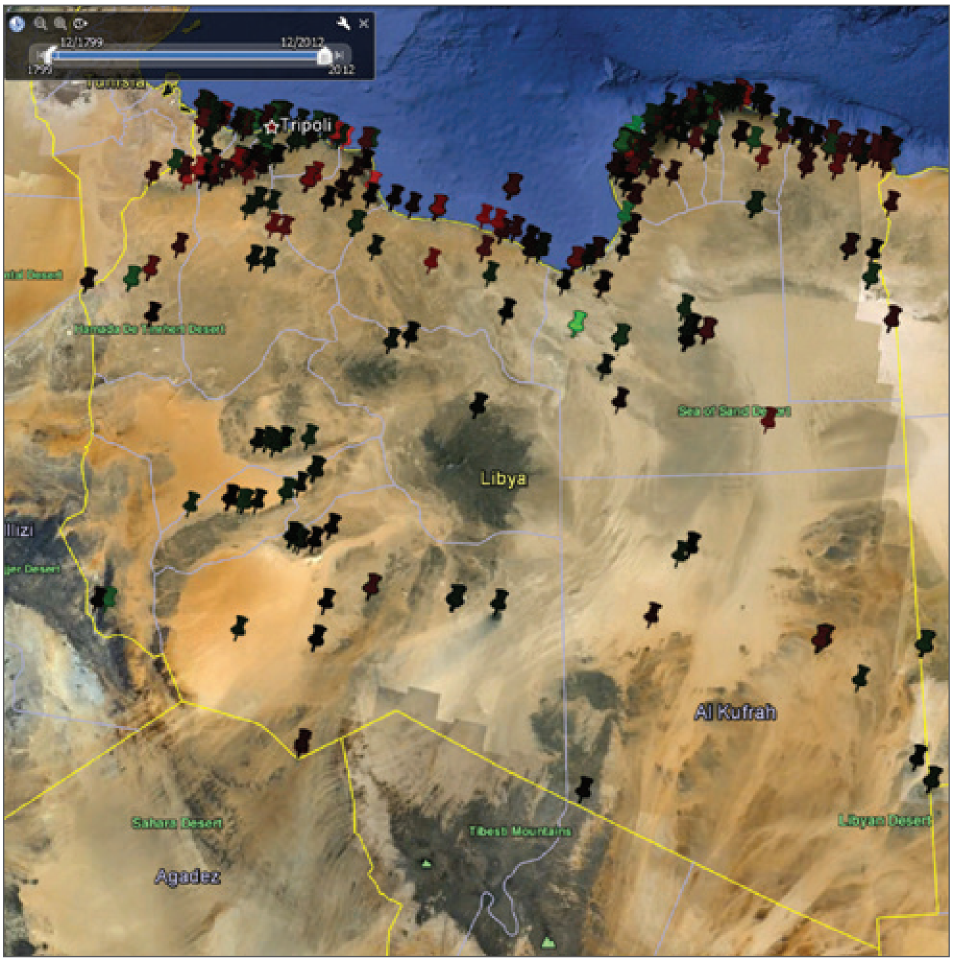


Figure 8: Interactive Google Earth file for Libya (see http://www.sgi.com/go/wikipedia/LIBYA-1800-2012.KML).

The one-way nature of Wikipedia

The Google Earth demonstration illustrates several limitations of Wikipedia's reliance on human editors to provide links between articles. For example, the Google Earth display shows mentions of Tajarhi, Libya in 1846 and 1848, reflecting that the entry for that city says slave trade traffic increased through there after Tunisia and Algeria abolished the trade, and also shows a mention in 1819 to reflect a description of it that year by the British naval explorer George Lyon (4). The article mentions both Tunisia and Algeria with respect to the slave trade, but those mentions are not links to those articles. The mention of George Lyon is also problematic, in that the actual Wikipedia page on his life is titled with his full name, George Francis Lyon" (5) and makes no mention of Tajarhi, only Tripoli and Murzuk, and is not linked from the Tajarhi page, requiring a visitor to manually keyword search on his name. The fact that these mentions of Tunisia, Algeria, and George Lyon have not been made into hyperlinks to those respective pages may at first seem to be only a small inconvenience. However, a data mining analysis of Wikipedia that looked only at which pages linked to which other pages (which is one of the most common ways Wikipedia is analyzed) would miss these connections. This illustrates the limitations of using linking data or other metadata to explore a large text corpus and the importance of examining the content itself.

Along those same lines are Wikipedia's "Infoboxes" in which human editors can create a table that appears in the sidebar of an article with important key facts about that article. These are often used as metadata to assign dates and locations to articles in data mining applications. For example, the American Civil War entry (6) has an Infobox with a rich assortment of details, including the locations and dates of the war. However, many articles do not contain such Infoboxes, even when the article focuses on a specific event. For example, the Barasa-Ubaidat War (7) between 1860-1890 in North-Eastern Libya, which started a year prior to the American Civil War, does not have an Infobox and the only information on the dates and locations of the conflict appear in the article text itself. The limitations of Infoboxes are something to keep in mind, as many studies and datasets make use of them as a machine-friendly proxy for the factual contents of Wikipedia (8).

Another trend in Wikipedia apparent in this Google Earth display is the tendency for a connection between two people or places to be mentioned in one of their respective entries, but not in the other's. For example, the entry for Tazirbu, Libya (9) notes that Gerhard Rohlfs was the first European to visit the oasis, in 1879. Rohlfs' own entry (10), however, notes only that in 1874 he embarked upon a journey to the Kufra basin in the same Kufra district in which Tazirbu is located, but does not mention Tazirbu itself or his visit there in 1879. The Kufra basin entry (11) notes that Rohlfs reached it in 1879, but again mentions nothing of Tazirbu or other details. The entry for Kufra District (12) in which both are located, mentions only that the name Kufra is a derivation of the Arabic word for a non-Muslim and cites one of Rohlfs' books, but does so only in the references list, and makes no mention of his travels in the text itself. Of course, Wikipedia entries must balance the desire to provide cross-links and updated information without turning each entry into a sea of links and repeated information. This is one of the areas where Wikipedia's openness really shines, in that it opens the door for computer scientists, interface designers, and others to apply data mining algorithms to develop new interfaces to Wikipedia and find new ways of finding and displaying these connections transparently.

The ability to display information from across Wikipedia temporally and spatially allows a reader to place a given event in the context of world events of the time period. For example, the Google Earth display contains a reference to Tripoli with respect to 1878 (the year prior to Rohlfs' visit to Tazirbu) to the entry for the Italo-Turkish War (13). At first glance this war appears to have no relation to 1879, having occurred 1911-1912. Yet, the opening sentence of the introductory paragraph notes that the origins of this war, in which Italy was eventually awarded the region of modern-day Libya, began with the Congress of Berlin in 1878. Thus, while likely entirely unrelated to Rohlfs' journey, it provides an additional point of context that can be found simply by connecting all of Wikipedia's articles together.

Thus, a tremendous amount of information in Wikipedia is one-way: one entry provides information about the connections between other entries, but those entries do not in turn mention this connection. If one was interested in the travels of Gerhard Rohlfs, a natural start would be to pull up his Wikipedia entry. Yet, his entry mentions only a brief synopsis of his African journey, with no details about the cities he visited. Even Paul Friedrich August Ascherson, who accompanied him on his journey, is not mentioned, while Ascherson's entry (14) prominently mentions his accompanying Rohlfs on the journey. One would have to keyword search all of Wikipedia for any mention of Rohlfs' name and then manually read through all of the material and synthesize their information in time and space to fully map out his journey. Using computational analysis, machines can do most of this work, presenting just the final analysis. This is one of the basic applications of data mining unstructured text repositories: converting their masses of words into knowledge graphs that recover these connections. In fact, this is what historical research is about: weaving a web of connections among people, places, and activities based on the incomplete and one-way records scattered across a vast archive of material.

The networks of Wikipedia

As a final set of analyses, four network visualizations were constructed to look at the broader structure of connections captured in Wikipedia.

Figure 9 shows how category tags are connected through co-occurrences in category-tagged articles. Wikipedia allows contributors to assign metadata tags to each article that describes the primary categories relevant to it. In this case, each category tag applied to an article was cross-linked with each other category tag for that article, across the entirety of Wikipedia, resulting in a massive network capturing how categories co-occur. This diagram illustrates a central core of categories around which other sub clusters of categories are tightly connected.

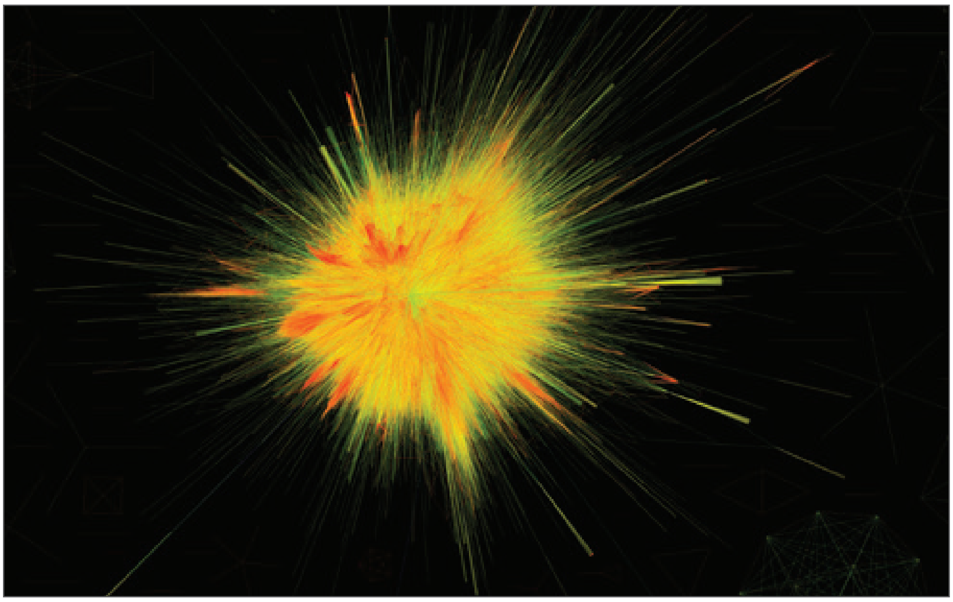


Figure 9: Network of co-occurrences of category tags across Wikipedia.

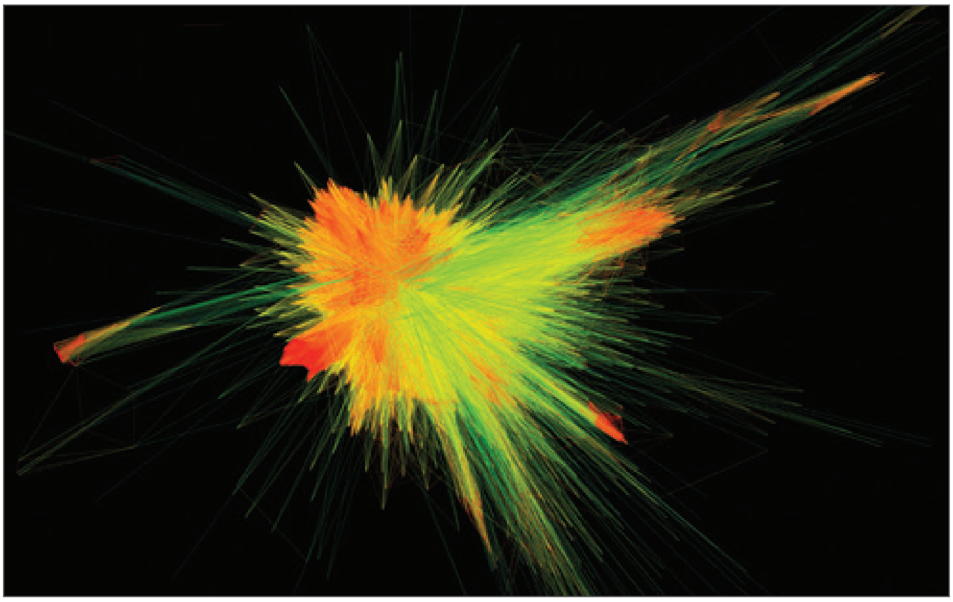


Figure 10: Network of co-occurrences of person names across Wikipedia.

Figure 10 shows the network of co-mentions of all person names across Wikipedia. In this case, a list of all person names appearing on each page was compiled and links formed to connect all person names appearing together in an article. This network shows a very different structure, which is far more diffuse with far greater clustering of small groups of people together.

Figure 11 shows the same approach applied to names of organizations. In this case, it is more similar to category tags, but shows more complex structure at the core, of clusters of names to which other clusters are tightly connected. Finally, Figure 12 shows the network of co-mentions of years across Wikipedia. This network illustrates that the closer to the present, the more Wikipedia content revolves around that year. This captures the fact that entries across Wikipedia tend to be updated with new information and events from the current year, which draws a connection between those earlier years and the present.

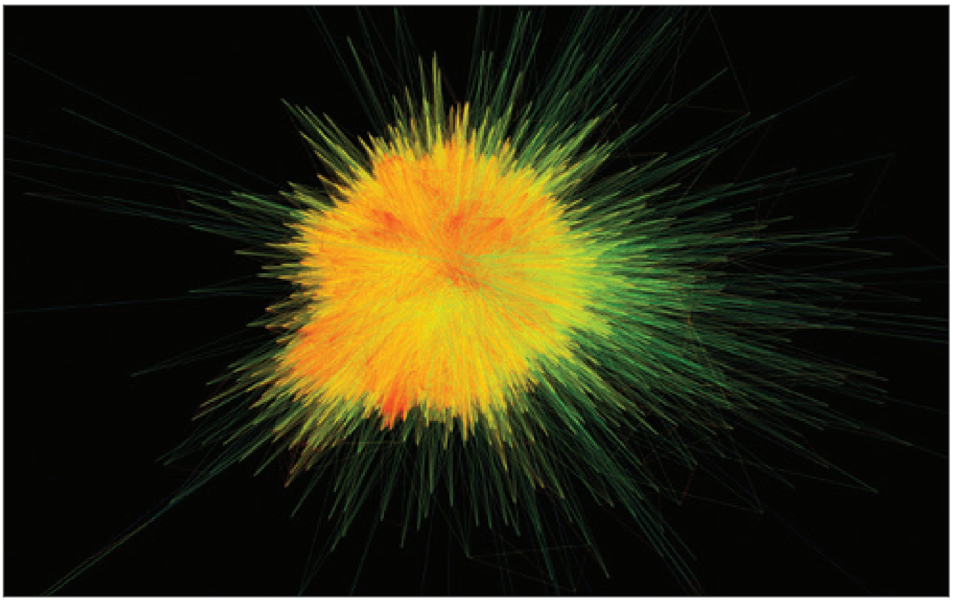


Figure 11: Network of co-occurrences of organization names across Wikipedia.

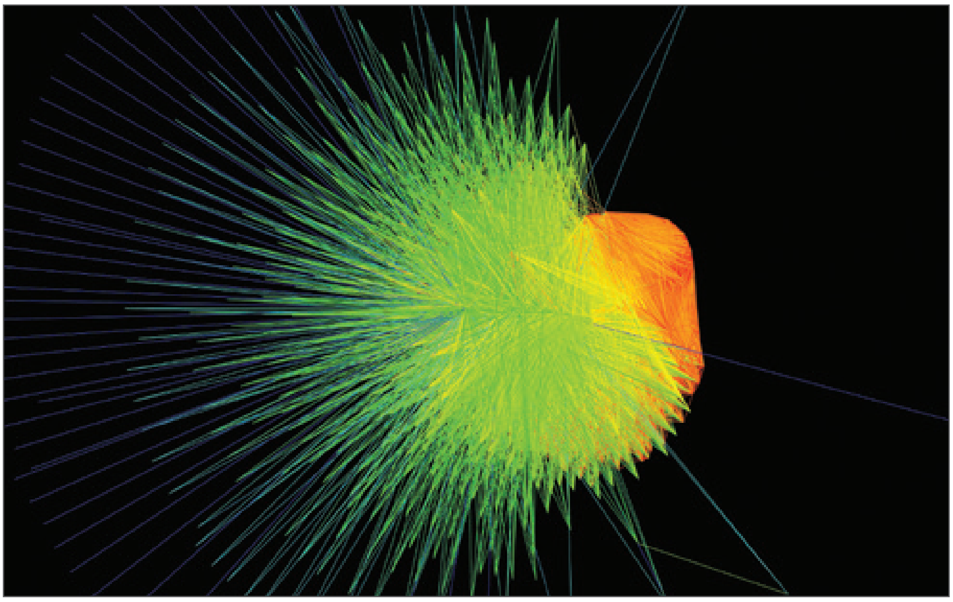


Figure 12: Network of co-occurrences of years across Wikipedia.

Conclusions

This study has surveyed the current landscape of the Big Data Humanities, Arts, and Social Sciences (HASS) disciplines and introduced the workflows, challenges, and opportunities of this emerging field. As emerging HASS scholarship increasingly moves towards data-driven computationally-assisted exploration, new analytical mindsets are developing around whole-corpus data mining, data movement, and metadata construction. Interactive exploration, visualization, and ad-hoc hypothesis testing play key roles in this new form of analysis, placing unique requirements on the underlying data storage and computation approaches. An exploration of Wikipedia illustrates all of these components operating together to visualize Wikipedia's view of world history over the last two centuries through the lens of space, time, and emotion.

Acknowledgements

The author wishes to thank Silicon Graphics International (SGI) for providing access to one of their UV2000 supercomputers to support this project.

References and Useful Links:

1. http://www.theatlantic.com/technology/archive/2012/03/the-missing-20th-century-how-copyright-protection-makes-books-vanish/255282/

2. Leetaru, K. (forthcoming). "Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards a Geographically Enriched Wikipedia", D-Lib Magazine.

3. Requires a free download of Google Earth http://www.google.com/earth/index.html

4. http://en.wikipedia.org/wiki/Tajarhi

5. http://en.wikipedia.org/wiki/George_Francis_Lyon

6. http://en.wikipedia.org/wiki/American_Civil_War

7. http://en.wikipedia.org/wiki/Barasa%E2%80%93Ubaidat_War

8. http://www.infochimps.com/collections/wikipedia-infoboxes

9. http://en.wikipedia.org/wiki/Tazirbu

10. http://en.wikipedia.org/wiki/Friedrich_Gerhard_Rohlfs

11. http://en.wikipedia.org/wiki/Kufra

12. http://en.wikipedia.org/wiki/Kufra_District

13. http://en.wikipedia.org/wiki/Italo-Turkish_War

14. http://en.wikipedia.org/wiki/Paul_Friedrich_August_Ascherson

Section 7:
The use of Big Datasets in bibliometric research

Henk F. Moed
Senior Scientific Advisor, Elsevier,
Amsterdam, The Netherlands

Introduction

Due to the increasing importance of scientific research for economic progress and competitiveness, and to new developments in information and communication technologies (ICT), the fields of bibliometrics and research assessment are rapidly developing. A few major trends can be identified:

- An increase in actual use of bibliometric data and indicators in research assessment;
- A strong proliferation of bibliometric databases and data-analytical tools; for instance, in the emergence of a range of journal subject classification systems and key words mapping tools;
- Indicators are becoming more and more sophisticated and fit-to-purpose; new approaches reveal that bibliometrics concerns much more than assessing individuals on the basis of journal impact factors;
- There is an increasing interest in measuring the effects of the use of bibliometric indicators upon the behavior of researchers, journal editors and publishers;
- Researchers, research evaluators and policy officials place an emphasis on the societal impact of research, such as its technological value or its contribution to the enlightenment of the general public;
- Last but not least, more and more projects aim to create and analyze large datasets by combining multiple datasets.

This article deals with the last trend mentioned and focuses on demonstrating which datasets are currently being combined by research groups in the field. It also discusses the aspects and research questions that could be answered using these large datasets. An overview is given in Table 1.

Examples

Downloads versus citations

For a definition of "usage" or "downloads" analysis and its context the reader is referred to a previous RT article on this topic (1). Figure 1 relates to journals included in ScienceDirect, Elsevier's full text article database. For each journal the average citation impact per article was calculated (generated in the third year after publication date), as well as the average number of downloads in full text format per article (carried out in the year of publication of the articles). Journals were grouped into disciplines; the horizontal axis indicates the number of journals in a discipline. In each discipline the Pearson correlation coefficient between a journal's downloads and its citations was calculated, and plotted on the vertical axis.

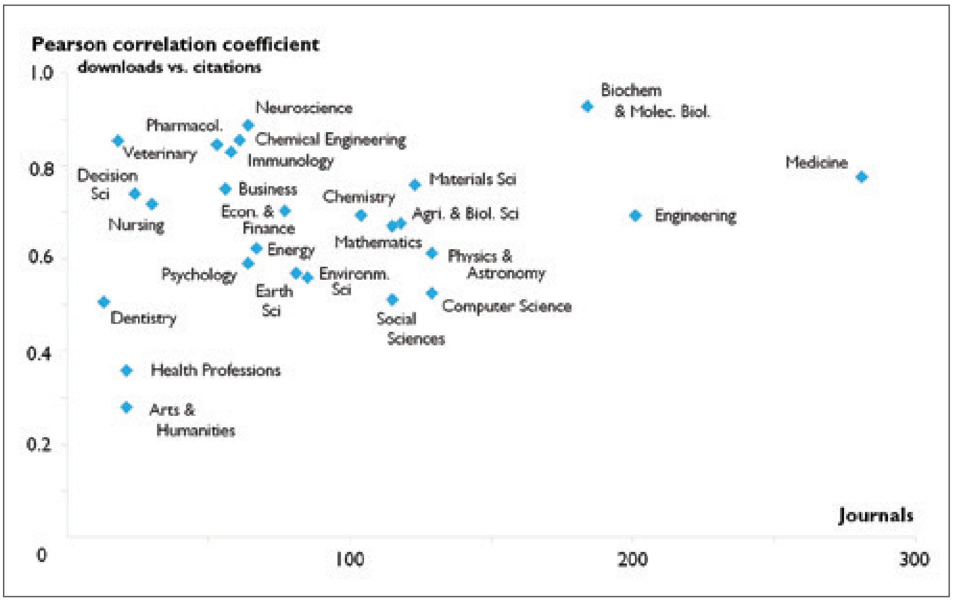


Figure 1: Downloads versus citations for journals in ScienceDirect.

Combined datasets	Studied phenomena	Typical research questions
Citation indexes and usage log files of full text publication archives	Downloads versus citations; distinct phases in the process of processing scientific information	What do downloads of full text articles measure? To what extent do downloads and citations correlate?
Citation indexes and patent databases	Linkages between science and technology (the science–technology interface)	What is the technological impact of a scientific research finding or field?
Citation indexes and scholarly book indexes	The role of books in scholarly communication; research productivity taking scholarly book output into account	How important are books in the various scientific disciplines, how do journals and books interrelate, and what are the most important books publishers?
Citation indexes (or publication databases) and OECD national statistics	Research input or capacity; evolution of the number of active researchers in a country and the phase of their career	How many researchers enter and/or move out of a national research system in a particular year?
Citation indexes and full text article databases	The context of citations; sentiment analysis of the scientific-scholarly literature	In what ways can one objectively characterize citation contexts? And identify implicit citations to documents or concepts?

Table 1: Compound Big Datasets and their objects of study.

Figure 1 reveals large differences in the degree of correlation between downloads and citations between disciplines. For instance, in Biochemistry and Molecular Biology the correlation is above 0.9, whereas in Dentistry, Social sciences, Health Professions, Arts and Humanities it is equal to or less than 0.5.

The interpretation of these findings is somewhat unclear. One hypothesis is based on the distinction between authors and readers. In highly specialized subject fields these populations largely overlap, whereas in fields with a more direct societal impact, the readers' population may consist mainly of professionals or even the general public who do not regularly publish articles. The hypothesis proposes that in the latter type of fields the correlation between downloads and citations is lower than in the first. Additional research, also conducted at the level of individual articles, is needed to further examine this hypothesis.

Patents and scientific articles

Earlier this year, Research Trends also published an article analyzing patent citations to journal articles, in order to measure the technological impact of research (2). The analysis focused on a subject field in the social sciences. It examined the characteristics of research articles published in Library Science journals and the manner by which they are cited in patents. Library science articles were found to be well cited in patents. The articles cited feature information retrieval and indexing, and information and documents management systems which pertain to electronic and digital libraries development. The citing patents focus on electronic information administration, navigation, and products and services management in commercial systems. Interestingly, the time span between the scientific invention and its use in technology may be up to 10 years. This finding illustrates the time delays one has to take into account when trying to measure technological or societal impact of scientific research. For an overview of this way of using patent citations, see (3).

Scopus author data versus OECD "input" statistics

Scopus, Elsevier's scientific literature database, containing meta-data of scientific publications published by more than 5,000 publishers in 18,000 titles, has implemented unique features that enable one to obtain an estimate of the number of active – i.e., publishing – authors in a particular year, country, and/or research domain, and also to track the "institutional" career of a researcher, providing information on the institutions in which a researcher has worked during his or her career. Research Trends issues 26 and 27 contained two articles by Andrew Plume presenting a first analysis of migration or brain circulation patterns in the database (4) (5).

Data accuracy and validation is also a relevant issue in this case. One way to validate author data is by comparing outcomes per country with statistics on the number of full time equivalents spent on research in the various institutional sectors, obtained from questionnaires and published by the OECD.

Country	Germany	UK	Italy	The Netherlands
OECD number of FTE Research 2007 (all sectors)	290,800	254,600	93,000	49,700
OECD number of FTE Research 2007 (Higher Education & Government sector)	116,600	159,100	56,200	23,800
Number of Publishing authors in Scopus	150,400	154,600	113,100	46,300
Ratio number of authors / Number of FTE Research (all Sectors)	0.52	0.61	1.22	0.93
Ratio number of authors / Number of FTE Research (Higher Education & Government sector)	1.29	0.97	2.01	1.95

Table 2: OECD and Scopus based "input" statistics for 4 European countries.

Table 2 presents statistics for 4 countries. Rather than comparing absolute numbers, it is interesting to calculate the ratios in the last two rows of the table. It is striking that these ratios differ substantially between countries. They are much higher for the Netherlands and Italy than they are for Germany and UK. This outcome points first of all towards the need to further validate Scopus-based numbers of active researchers. On the other hand, it also raises the question whether the various countries have applied the same definition of FTE research time in their surveys.

Books and journals

Scientific-scholarly books are generally considered as important written communication media, especially in social sciences and humanities. There is an increasing interest in studies of the function and quality of books and book publishers in the various domains of science and human scholarship. Thomson Reuters has launched its Book Citation Index. The Google Books project aims to digitalize millions of books, including many scientific-scholarly ones. Expanding a primarily journal-based citation index with scholarly book sources has two advantages. Not only is the set of source publications expanded with relevant sources, but the enormous reservoir of cited references given in journal articles to book items is used more efficiently.

Citations and full texts

The availability of full text research articles in electronic format gives us the opportunity to conduct textual analyses of all of an article's content – not just the meta-data extracted by indexing databases. The citation contexts can be analyzed linguistically, and sentiment analyses can be conducted to reveal how the citing author appreciates a cited work. Henry Small and Richard Klavans used citation context analysis as an additional tool for the identification of scientific breakthroughs (6). In one of its next issues Research Trends will publish an article on a detailed citation context analysis in one particular journal focusing on cross-disciplinary citations.

Concluding remarks

The overview above is not complete, and many important contributions to the analysis of big, compound bibliometric datasets were not mentioned in this paper. But the examples presented above illustrate the theoretical and practical relevance of combining bibliometric, or, more generally, statistical datasets, show how this can be done, and indicate which issues a big, compound, bibliometric dataset enables us to address.

References:

1. Lendi, S. & Huggett, S. (2012) "Usage: an alternative way to evaluate research", Research Trends, No. 28 http://www.researchtrends.com/issue28-may-2012/usage-an-alternative-way-to-evaluate-research/

2. Halevi, G. & Moed, H.F. (2012) "Patenting Library Science Research Assets", Research Trends, No. 27 http://www.researchtrends.com/issue-27-march-2012/patenting-library-science-research-assets/

3. Breschi, S. & Lissoni, F. (2004) "Knowledge Networks from Patent Data. In: Moed, H.F., Glänzel, W., and Schmoch, U. (eds.). Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems. Dordrecht (the Netherlands): Kluwer Academic Publishers, 613-644.

4. Plume, A. (2012) "The evolution of brain drain and its measurement: Part I", Research Trends, No. 26 http://www.researchtrends.com/issue26-january-2012/the-evolution-of-brain-drain-and-its-measurement-part-i/

5. Plume, A. (2012) "The evolution of brain drain and its measurement: Part II", Research Trends, No. 27 http://www.researchtrends.com/issue-27-march-2012/the-evolution-of-brain-drain-and-its-measurement-part-ii/

6. Small, H. & Klavans R. (2011). "Identifying Scientific Breakthroughs by Combining Co-citation Analysis and Citation Context". Paper presented at the Proceedings of 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2011).

Section 8:

Did you know?

…It is our birthday

Happy Birthday Research Trends!

On Friday the 28th September it is exactly five years to the day since Research Trends sent its first issue out.

We have been cited by highbrow media outlets such as Nature Magazine and Die Zeit. And we have been cited in the published literature:

- Arencibia-Jorge, R., & de Moya-Anegón, F. (2010) "Challenges in the study of Cuban scientific output", Scientometrics, 83(3), 723–737.
- Kähler, O. (2010) "Combining peer review and metrics to assess journals for inclusion in Scopus", Learned Publishing, 23(4), 336–346.
- Leydesdorff, L. (2012) "World shares of publications of the USA, EU-27, and china compared and predicted using the new Web of Science interface versus Scopus", Profesional De La Informacion, 21(1), 43–49.
- Miguel, S., Chinchilla-Rodriguez, Z., & de Moya-Anegón, F. (2011) "Open access and Scopus: A new approach to scientific visibility from the standpoint of access", Journal of the American Society for Information Science and Technology, 62(6), 1130–1145.
- Vanclay, J. K., & Bornmann, L. (2012) "Metrics to evaluate research performance in academic institutions: A critique of ERA 2010 as applied in forestry and the indirect H2 index as a possible alternative", Scientometrics, 91(3), 751–771.
- Kirby, A. (2012) "Scientific communication, open access, and the publishing industry", Political Geography, 31(5), 256–259.
- Jacsó, P. (2012) "Grim tales about the impact factor and the h-index in the Web of Science and the Journal Citation Reports databases: Reflections on Vanclay's criticism", Scientometrics, 92(2), 325–354.
- Moed, H.F. et al. (2012) "Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way", Scientometrics, 92(2), 367–376.

One of our most popular issues was our special issue on de Solla Price, issue 7, 2008.

Research Trends: Editorial Board

You can find more information on www.researchtrends.com or contact us at researchtrends@elsevier.com

Notes:

ELSEVIER