Big Data, Digital Humanities and the New Knowledge Environments of the 21st century

Chad Gaffield, Ph.D FRSC
President
Social Sciences and Humanities Research Council of Canada

May 15, 2013
Social sciences and humanities research builds knowledge about people in the past and present, with a view toward creating a better future.
<table>
<thead>
<tr>
<th>TALENT</th>
<th>INSIGHT</th>
<th>CONNECTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supporting students and postdoctoral fellows to develop next generation researchers and leaders across society</td>
<td>Supporting excellent research to advance knowledge and build understanding about people, communities and societies</td>
<td>Supporting the exchange of knowledge to maximize the intellectual, cultural, social and economic impacts of social sciences and humanities research</td>
</tr>
</tbody>
</table>

**PARTNERSHIPS** - Within the academic community and between academia, industry, government, not-for-profits and communities.
Technology-driven age?

Rather, new thinking and behaviour are being enabled, accelerated and influenced in iterative ways by digital technologies.
“I'm afraid you have humans.”
Innovative businesses and organizations now seek to be customer-focused in the marketplace, user-oriented in service industries, partner-driven in collaboration, employee-empowered in workplaces, citizen-engaged in politics, student-centered in schools, and patient-focused in health.
Nearly every company is getting more data about customers, sales and interactions than it can quickly put to good use. One fascinating solution is the development of artificial intelligence platforms that can extract insight from large datasets and communicate those insights in narrative form.

http://mitchspeers.com/2013/03/26/big-data-for-humans/
Nearly every school is getting more data about students, learning and engagement than it can quickly put to good use. One fascinating solution is the development of artificial intelligence platforms that can extract insight from large datasets and communicate those insights in narrative form.

With apologies to http://mitchspeers.com/2013/03/26/big-data-for-humans/
Nearly every hospital is getting more data about patients, interventions, and outcomes than it can quickly put to good use. One fascinating solution is the development of artificial intelligence platforms that can extract insight from large datasets and communicate those insights in narrative form.

With apologies to http://mitchspeers.com/2013/03/26/big-data-for-humans/
Nearly every government is getting more data about citizens, user needs and desires, and service standards than it can quickly put to good use. One fascinating solution is the development of artificial intelligence platforms that can extract insight from large datasets and communicate those insights in narrative form.

With apologies to http://mitchspeers.com/2013/05/20/big-data-for-humans/
DATA TSUNAMI
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
ENTENDA O QUE É BIG DATA
O MEGAFENÔMENO DIGITAL QUE TRANSFORMA EM RIQUEZA DADOS PESSOAIS, POSTS, TUITES, E-MAILS E ATÉ CLIQUES
Big data spans four dimensions: Volume, Velocity, Variety, and Veracity.

source: IBM definition
Volume: Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information.

Turn 12 terabytes of Tweets created each day into improved product sentiment analysis. Convert 350 billion annual meter readings to better predict power consumption.
**Velocity:** Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

Scrutinize 5 million trade events created each day to identify potential fraud.
Analyze 500 million daily call detail records in real-time to predict customer churn faster.
**Variety**: Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.

Monitor 100’s of live video feeds from surveillance cameras to target points of interest. Exploit the 80% data growth in images, video and documents to improve customer satisfaction.
Veracity: 1 in 3 business leaders don’t trust the information they use to make decisions. How can you act upon information if you don’t trust it? Establishing trust in big data presents a huge challenge as the variety and number of sources grows.
People ≠ Particles
People ≠ Particles

Why not? unique individuals + specific context
People ≠ Particles

Why not? unique individuals + specific context
But – data analytics have lots to share
What is DIGITAL Humanities?

It's the use of the newest digital technologies to answer traditional questions in the Humanities: Did Shakespeare write all of his plays? What is the genetic code of Canadian culture? Who painted “The Parable of the Unfaithful Servant”? How would you read a million books?
“We have observed that topics are connected by a net of relationships that can be seen as a complex network.”

Visit [http://cultureplex.ca/](http://cultureplex.ca/) for more information.
This project will examine topic lifecycles across heterogeneous corpora, including not only scholarly journal articles, but also social media, news articles, and other forms of digital text. The goal is to develop new and accessible tools. This work should identify which scholarly activities are indicative of emerging areas and identify datasets that should no longer be marginalized, but built into understandings and measurements of scholarship.

While the growth of large-scale datasets has enabled examination within scientific datasets, there is little research that looks across datasets.
SALAMI

SALAMI (Structural Analysis of Large Amounts of Music Information) is an innovative and ambitious computational musicology project. To date, musical analysis has been conducted by individuals and on a small scale. Our computational approach, combined with the huge volume of data now available from such sources as the Internet Archive, will: a) deliver a very substantive corpus of musical analyses in a common framework for use by music scholars, students, and beyond; and, b) establish a methodology and tooling which will enable others to add to this in the future and to broaden the application of the techniques we establish. A resource of SALAMI’s magnitude empowers musicologists to approach their work in a new and different way, starting with the data, and to ask research questions that have not been possible before.

There are two resources available on this site:

- **Background.** The background page (consisting mainly of the proposal for the Digging Into Data grant that SALAMI was awarded in 2009) gives an thorough overview of the SALAMI project.
- **Annotation data.** Visit this page to access the annotation data and to learn about how it was collected.

Through a Digging Into Data grant, this research was supported by the Social Sciences and Humanities Research Council, by the National Science Foundation, and by JISC.
IMPACT’s Mission

The IMPACT Radiological Mummy Database is a large-scale, multi-institutional collaborative research project devoted to the scientific study of mummified remains, and the mummification traditions that produced them, through non-destructive medical imaging technologies.

IMPACT focuses on the body, made artifact through cultural or natural intervention, in bioarchaeology, epidemiology, and social archaeology studies of past human societies and their genetic and cultural descendants.

News

The IMPACT Project has been successful in the 2011 Digiing into Data Challenge, and will be receiving support from SSHRC and NEH in 2012 & 2013!

IMPACT has made arrangements with 49 institutions, to date, to share their radiographic mummy data, with live online access expected in Summer 2012.

EN FR ES DE العربية 中文
IP Policy Contact Site Map

Designed by A.Wade. Last updated June 2012.

Visit http://impactdb.uwo.ca/IMPACTdb/Index.html for more information.
What are the intersections between biomedicine and humanities scholarship? How might biomedical research methodologies influence humanities inquiry? What are the intersections between biomedicine and humanities scholarship? How might biomedical research methodologies influence humanities inquiry?

Wednesday, April 10, 2013

Symposium:
Thursday, April 11 to Friday, April 12, 2013

APPLY NOW

Check out our latest news:
Keynote Lecturer: David B. Sears
The MITH Blog
The NEH Press Release
The NLM Press Release

Brett Bobley’s recent blog on how Shared Horizons came to be.
NLM & National Endowment for the Humanities to Cooperate on Initiatives of Common Interest

The National Library of Medicine (NLM), the world’s largest medical library and a component of the National Institutes of Health, and the National Endowment for the Humanities (NEH) are forming a new partnership to develop initiatives that bring together scholars, scientists, librarians, doctors and cultural heritage professionals from the humanities and biomedical communities in order to share expertise and develop new research, teaching and learning by the humanities and biomedical communities.

As initial efforts, the partners will work together to:

- Explore areas of mutual interest for research, particularly in the fields of digital humanities and the history of medicine;
- Develop and participate in curricula and courses, training and internship opportunities, and other educational initiatives; and
- Develop initiatives to increase access to careers in medicine and health, with a particular interest in reaching students who are under-represented in the fields.

About the partner institutions:

The National Endowment for the Humanities is an executive-branch, independent grant-making agency of the United States of America dedicated to supporting research, education, preservation, and public programs in the humanities and in those social sciences that use humanistic methods. NEH accomplishes this mission by providing grants for high-quality humanities projects to cultural institutions, such as museums, archives, universities, public television and radio stations, and to individual scholars.

Since its founding in 1836, the National Library of Medicine has played a pivotal role in translating biomedical research into practice. NLM, part of the National Institutes of Health, is the world’s largest medical library and an essential information resource to many millions of items in its collection. A leader in information innovation, it is the developer of electronic information services used by scientists, health professionals and the public around the world. NLM makes its information available with the help of the National Network of Libraries of Medicine, which consists of 5,600 member institutions, including eight Regional Medical Libraries. NLM conducts and supports research that applies science to meet the information needs of clinicians, public health administrators, biomedical researchers and consumers.

Big data meets language analytics: Mapping the evolution of culture

Google Books Ngram Reader – bigram results of search for terms “nationalize” and “privatize” across database of 15 million published sources.

Courtesy SSHRC grantee Ian Milligan (www.ianmilligan.ca)
Big data meets language analytics: Mapping the evolution of culture

Google Books Ngram Reader – unigram result of search for “Canada” across database of 15 million published sources.

Courtesy SSHRC grantee Ian Milligan (www.ianmilligan.ca)
Big data meets language analytics: Mapping the evolution of culture

Google Books Ngram Reader – search for reference to Canada’s 5 largest cities across database of 15 million published sources.

Courtesy SSHRC grantee Ian Milligan (www.ianmilligan.ca)
Computing in the Humanities

Proceedings of the Third International Conference on Computing in the Humanities

Sponsored by the University of Montreal and the University of Waterloo
August 1-4, 1977
at Waterloo, Ontario

Editors:
Serge Layeghe and John S. North
The Emergence of Digital SSH Scholarship in Canada

Examples:

- **1978**: Canadian Committee on History and Computing

- **1986**: COCH/COSH (Consortium for Computing in the Humanities) which became SDH/SEMI (Society for Digital Humanities/Société pour l'étude des médias interactifs).
New Data for Understanding the Human Condition

International Perspectives

OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences

February 2013
``Data-driven and evidence-based research is fundamental to understanding and responding effectively and efficiently to global challenges related to the health and wellbeing of populations around the world.``

OECD Global Science Forum Report Feb 2013
``Spurred by the rapid growth in new forms of data collected in conjunction with commercial transactions, internet searches, social networking, and the like, and by technological advances in the capacity to access and link existing survey, census, and administrative data sets, the potential payoff for international and multidisciplinary collaboration of scientific groups to address these challenges is increasing rapidly.``
OECD Global Science Forum Report Feb 2013
Chair
United States of America: Barbara Entwisle
Vice Chancellor for Research
University of North Carolina, Chapel Hill

Vice Chair
United Kingdom: Peter Elias
Institute for Employment Research
University of Warwick

Other members included:

Brazil: José Eduardo Cassiolato
(from June 2011)
Economics Institute
Federal University of Rio de Janeiro

Canada: Chuck Humphrey
University of Alberta and representing the
Social Sciences and Humanities Research
Council of Canada
Recommendation 1: national research funding agencies should collaborate internationally to provide resources for researchers to assess the research potential and to develop new methods to understand the opportunities and limitations offered by new forms of data to address important research areas.

OECD Global Science Report 2013
International Forum of Funding Agencies, Founding meeting, Ottawa Canada, Sept 2007
Digging into Data
First launch, Ottawa Canada, 2009
Transatlantic Platform
for the Social Sciences and Humanities

Management and administrative coordination (WP 1)

Strong Trans-Atlantic partnership (WP 2)
- Enhancing collaboration between partners
- Stimulating networking, ongoing research
- Communicating best practices

Management and administrative coordination (WP 1)

Digital Pilot Project (WP 3)
- Digital pilot project

Management and administrative coordination (WP 1)

Thematic challenges (WP 4)
- Scoping exercises
- Recommendations for SSH opportunities
“It’s in Apple’s DNA that technology alone is not enough. It’s technology married with Liberal Arts, with the Humanities, that yields the results that make our hearts sing.”

Steve Jobs, 1955-2011
Percentage of students identifying Apple or Google in their top five preferred places to work

Source: Globecampus.ca
"We are going through a period of unbelievable growth and will be hiring about 6,000 people this year - and probably 4,000-5,000 from the humanities or liberal arts."

- Marissa Mayer, Vice-President of Consumer Products, Google
HOW THE DIGITAL TRIANGLE IS FRAMING THE 21ST CENTURY

Digital Technologies  

Digital Content

Digital Literacies
The Automated Archivist: 
Interdisciplinarity and the Process of Historical Research 

CHAD GAFFIELD & PETER BASKERVILLE

The basis of most historical research including social science history is quite unsystematic. This characteristic results from the ways in which researchers find and choose historical sources for examination. Despite claims to be systematic, historians still tend to identify relevant evidence in impressionistic ways. Many social science histories involve the rigorous study of a source happily discovered by chance. Of course, access to the past has never been easy. Researchers have always lamented a presumed lack of "essential" records. Nonetheless, the actual ways we discover existing evidence have received little attention despite the fact that this process is fraught with difficulties and hidden dangers especially for researchers of a social scientific bent. Do not the presuppositions of social science history extend to the identification of sources? How do we know when we have all the "relevant data" for a particular project? Can systematic data analysis be justifiably built upon unsystematic identification of sources?

Our consideration of such questions is a central part of the Vancouver Island Project (Baskerville and Gaffield, 1983–84), which is preparing a machine-readable research tool for all local public repositories on British Columbia's Vancouver Island (total population c. 500,000). The main group of repositories are related

Chad Gaffield and Peter Baskerville both teach in the History Department of the University of Victoria in British Columbia.

“New information technologies represent one of the major impacts on humanities teaching and research. They also present an exciting opportunity for scholars, teachers and students to become informed partners and innovators.”

To examine and interpret individuals and their cultures, researchers currently use three fundamental kinds of digital information: images, text and sound.

These digital forms of information are, however, very sensitive to changes in the technologies through which they are created, analyzed, published and preserved. In recent decades, innovative technologies have transformed the very definition of text and its relationship to image and sound. To benefit fully from these new technologies, researchers must not only be aware of technological developments, but also be directly involved in them.
The overall objectives of the ITST program are to:

- reflect on, interpret, and analyze new digital media, multimedia, and text-based computing technologies, and integrate these into humanities and social sciences research;

- bring together theorists, experimentalists, and technologists from different disciplines to share and nurture ideas and methods that challenge research to advance through the use of audio-visual and text-based technologies; and

- facilitate the creation of national and international networks of, and partnerships among, researchers, industries, governments, and individuals that will promote and sustain social sciences and humanities research and resources worldwide.
Possible topics and areas to be addressed through ITST support include:

- electronic editing and publishing;
- e-literature;
- Web programming;
- immersive and virtual environments in multimedia research;
- textual analysis;
- 3D imaging technology;
- creativity, culture and computing;
- digital image design;
- information aesthetics;
- computer gaming; and
- knowledge transfer of research results to fellow researchers, decision-makers and the public at large.
<table>
<thead>
<tr>
<th>Citizenship, Nationality and Religion</th>
<th>Profession, Occupation, Trade or Means of Living</th>
<th>Wage Earned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country or place of birth (if in Canada, specify province or territory)</td>
<td>Chief occupation or trade</td>
<td>Employees working or earning</td>
</tr>
<tr>
<td>Year of enumeration or immigration</td>
<td>Employment other than chief occupation or trade, if any.</td>
<td>Wages in 1900 or 1910, if any.</td>
</tr>
<tr>
<td>Year of arrival as immigrant</td>
<td>Nationality</td>
<td>Employees employed in 1900 or 1910, if any.</td>
</tr>
<tr>
<td>Race or tribal origin</td>
<td>Religion</td>
<td>Employees employed in other than chief occupation or trade, if any.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citoyenneté, nationalité et religion</td>
<td>Profession, emploi, métier ou moyen d'existence</td>
<td>Emploi</td>
</tr>
<tr>
<td>Pays ou lieu de naissance (si lieu de naissance est en Canada, spécifier la province ou le territoire)</td>
<td>Emploi principal ou métier.</td>
<td>Travail à son compte (voir instructions).</td>
</tr>
<tr>
<td></td>
<td>Origine selon la race ou la tribu</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nationalité</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Religion</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>23</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Count</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Czech &amp; Slovak</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Ukrainian</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Native Indian</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Jewish</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

*If "W," "J," or "L" in 20, or "YES" in 21, ask Questions 23 to 27.*

Questions 23-25 and 27 refer to this person's employment during week ending June 2, 1951. If not working that week, these questions refer to most recent employment. If "L" in 20 and this person never worked, enter "NONE" in 25 and usual activity in 26.

**23. Name of Firm or Business** — For whom did this person work?

**27. Class of Worker**
A series of Nova Scotia CSD maps (1871–1951), reconstituted by using every place-name in each EA and CSD list; it is the best reconstitution available for those CSDs.

Enlarged view of A. H. Clark’s map for Pictou CD and CSDs. It was georeferenced and superimposed on 2001 Statistics Canada geofiles ...

from which we extracted most thin lines and created some others from hydro, roads, and railway geofiles (thicker lines) to reconstitute CCRI 1911 CSD polygons (conforming to the underlying map by A. H. Clark).

FIGURE 2. Processing 1911 Nova Scotia CSDs: The example of Pictou CD, 1911. *(Putting)* source. Andrew H. Clark. [AU: please provide date here for Clark collection that I will now include in the Reference]
FIGURE 8. CCRI database cycles.
CCRI was supported by:

Canada Foundation for Innovation,
Ontario Innovation Trust,
Le Ministère de l’Éducation du Québec,
The Harold Crabtree Foundation,
IBM Canada,
The International Microdata Access Group (IMAG),
L’Institut de la Statistique du Québec,
The National Archives of Canada,
The Newfoundland Statistics Agency,
Statistics Canada and other partners.
Inferring and Revising Theories with Confidence:
Analyzing the 1901 Canadian Census

Chris Drummond
Institute for Information Technology
National Research Council Canada
Ottawa, Ontario, Canada K1A 0R6

Stan Matwin
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5

Chad Gaffield
Institute of Canadian Studies
University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5
• Data are not `neutral,` not `objective`
• Data are not `neutral`, `not `objective`
• In creating data, humans make choices, decisions all along the way.
  – What data to collect?
  – How to collect it?
  – How to categorize the results?
  – How to interpret?
  – How to attribute meaning? relevance? Etc.
• Data do not `speak for themselves`
• The `facts` do not speak for themselves
• Turning data into insight depends upon human interpretation
• Human decisions embedded in software, algorithms
• When the data concerns people, all these decisions necessarily reflect `theories` (assumptions, etc) about human thought and behaviour
Since humans are diverse (unique individuals and specific context), households, communities, societies, cultures are diverse
Dr. Reut Shemesh

Want to Know My Future?

New genetic tests can point to risks—
but not always a cure

BY BONNIE ROCHMAN
The personalized medicine revolution is almost here

http://venturebeat.com/2013/01/27/the-personalized-medicine-revolution-is-almost-here/
Integrated Public Use Microdata Series, International

census microdata for social and economic research

IPUMS-International is a project dedicated to collecting and distributing census data from around the world. Its goals are to:

- Collect and preserve data and documentation
- Harmonize data
- Disseminate the data absolutely free!

68 countries - 211 censuses - 480 million person records
SSHRC drives Canada’s contribution to international research partnership initiative:

**ChartEx**
- exploring the full text content of digital historical records

**Digging into Human Rights Violations: Anaphora Resolution and Emergent Witnesses**
- Developing an automated reader for large text archives of human rights abuses.

**Electronic Locator of Vertical Interval Successions (ELVIS): The First Large Data-Driven Research Project on Musical Style**
- studying changes in Western musical style from 1300 to 1900

**An Epidemiology of Information: Data Mining the 1918 Influenza Pandemic**
- Harnessing the power of data mining techniques with the interpretive analytics of the humanities and social sciences to understand how newspapers shaped public opinion.

**IMPACT Radiological Mummy Database**
- providing mummy and medical researchers with a large-scale comparative database of medical imaging of mummified human remains.
IMPACT Radiological Mummy Database
• providing mummy and medical researchers with a large-scale comparative database of medical imaging of mummified human remains.

Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850-1911
• making use of novel data-mining technology to exploit one of the largest population databases in the world, a vast collection of harmonized 19th and early 20th century census microdata.

Structural Analysis of Large Amounts of Music Information (SALAMI)
• SALAMI (Structural Analysis of Large Amounts of Music Information) is an innovative and ambitious computational musicology project.

Harvesting Speech Datasets for Linguistic Research on the Web
• This project will harvest audio and transcribed data from podcasts, news broadcasts, public and educational lectures and other sources to create a massive corpus of speech.

Towards Dynamic Variorum Editions
• The creation of a framework to produce "dynamic variorum" editions of classics texts that enable the reader to automatically link not only to variant editions but also to relevant citations, quotations, people, and places that are found in a digital library of over one million primary and secondary source texts.
Further reading

Emerging research and commentary on culturomics

• Culturomics.org
• Kalev Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. First Monday, 17 August 2011, vol. 16 Number 9.
• Alasdair Wilkins. Cultural genome project mines Google Books for the secret history of humanity. io9, December 16, 2010
### Box 2

#### Forms of data with new research potential

<table>
<thead>
<tr>
<th>Broad category of data</th>
<th>Detailed categories</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category A: Government transactions</strong></td>
<td>Individual tax records</td>
<td>Income tax; tax credits</td>
</tr>
<tr>
<td></td>
<td>Corporate tax records</td>
<td>Corporation tax; sales; tax, value added tax</td>
</tr>
<tr>
<td></td>
<td>Property tax records</td>
<td>Tax on sales of property; tax on value of property</td>
</tr>
<tr>
<td></td>
<td>Social security payments</td>
<td>State pensions; hardship payments; unemployment benefits; child benefits</td>
</tr>
<tr>
<td></td>
<td>Import/export records</td>
<td>Border control records; import/export licensing records</td>
</tr>
<tr>
<td><strong>Category B: Government and other registration records</strong></td>
<td>Housing and land use registers</td>
<td>Registers of ownership</td>
</tr>
<tr>
<td></td>
<td>Educational registers</td>
<td>School inspections; pupil results</td>
</tr>
<tr>
<td></td>
<td>Criminal justice registers</td>
<td>Police records; court records</td>
</tr>
<tr>
<td></td>
<td>Social security registers</td>
<td>Registers of eligible persons</td>
</tr>
<tr>
<td></td>
<td>Electoral registers</td>
<td>Voter registration records</td>
</tr>
<tr>
<td></td>
<td>Employment registers</td>
<td>Employer census records; registers of persons joining/leaving employment</td>
</tr>
<tr>
<td></td>
<td>Population registers</td>
<td>Births; marriages; civil unions; deaths; immigration/emigration records; census records</td>
</tr>
<tr>
<td></td>
<td>Health system registers</td>
<td>Personal medical records; hospital records</td>
</tr>
<tr>
<td></td>
<td>Vehicle/driver registers</td>
<td>Driver licence registers; vehicle license registers</td>
</tr>
<tr>
<td></td>
<td>Membership registers</td>
<td>Political parties; charities; clubs</td>
</tr>
<tr>
<td><strong>Category C: Commercial transactions</strong></td>
<td>Store cards</td>
<td>Supermarket loyalty cards</td>
</tr>
<tr>
<td></td>
<td>Customer accounts</td>
<td>Utilities; financial institutions; mobile phone usage</td>
</tr>
<tr>
<td></td>
<td>Other customer records</td>
<td>Product purchases; service agreements</td>
</tr>
<tr>
<td><strong>Category D: Internet usage</strong></td>
<td>Search terms</td>
<td>Google©, Bing©, Yahoo© search activity</td>
</tr>
<tr>
<td></td>
<td>Website interactions</td>
<td>Visit statistics; user generated content</td>
</tr>
<tr>
<td></td>
<td>Downloads</td>
<td>Music; films; TV</td>
</tr>
<tr>
<td></td>
<td>Social networks</td>
<td>Facebook©, Twitter©, LinkedIn©</td>
</tr>
<tr>
<td></td>
<td>Blog; news sites</td>
<td>Reddit</td>
</tr>
<tr>
<td><strong>Category E: Tracking data</strong></td>
<td>CCTV images</td>
<td>Security/safety camera recordings</td>
</tr>
<tr>
<td></td>
<td>Traffic sensors</td>
<td>Vehicle tracking records; vehicle movement records</td>
</tr>
<tr>
<td></td>
<td>Mobile phone locations: GPS data</td>
<td></td>
</tr>
<tr>
<td><strong>Category F: Satellite and aerial imagery</strong></td>
<td>Visible light spectrum</td>
<td>Google Earth©</td>
</tr>
<tr>
<td></td>
<td>Night-time visible radiation</td>
<td>Landsat</td>
</tr>
<tr>
<td></td>
<td>Infrared, radar mapping</td>
<td></td>
</tr>
</tbody>
</table>
A DISCIPLINE GOES DIGITAL

The humanities mine cultural databases

The digital humanities — the use of algorithms to search for meaning in databases of text and other media — have been around for decades. Some trace the field’s origins to Roberto Busa, an Italian priest who, in the late 1940s, teamed up with IBM to produce a searchable index of the works of thirteenth-century theologian Thomas Aquinas.

But the field has taken on new life in recent years. Journals have sprouted up and professional societies are blooming. Some universities are now requiring graduate students in the humanities to take statistics and computer-science courses. Funding — far harder to come by in the humanities than in the sciences — flows slightly more generously to those willing to adopt the new methods. This year, the US National Endowment for the Humanities, in collaboration with the National Science Foundation and research institutions in Canada and Britain, plans to hand out 20 grants in the digital humanities, worth a total of US$6 million.

Many researchers in the digital humanities use textual databases composed primarily of books — as Erez Lieberman Aiden does in his ‘culturomics’ project (see ‘Heavy-duty data’). Franco Moretti, a literary scholar at Stanford University in California, has shown that genres of fiction — Gothic novels, for example, or romance — have a textual ‘fingerprint’ that is apparent even in simple frequency counts of nouns, verbs and prepositions. “These genres are different at every scale,” he says.

HEAVY-DUTY DATA

The computer-storage space required to support projects in the digital humanities is now starting to rival that of big-science projects.

BIG SCIENCE

SLOAN DIGITAL SKY SURVEY
50 TERABYTES
The survey, begun in 1998 using a 2.5-metre telescope in New Mexico, has discovered nearly half-a-billion asteroids, stars, galaxies and quasars.

GENBANK
530 GIGABYTES
This database, which stores publicly available sequenced DNA, included 127 billion bases at the latest count.
BIG HUMANITIES

CULTUROMICS N-GRAMS VIEWER
300 GIGABYTES (English only)
The string of letters in this corpus of 5 million books is 1,000 times longer than the human genome.

YEAR OF SPEECH
1 TERABYTE
This database includes recordings from telephone conversations, broadcast news, talk shows and US Supreme Court arguments.

UNIVERSITY OF SOUTHERN CALIFORNIA SHOAH ARCHIVE
200 TERABYTES
This archive stores 52,000 videotaped interviews with Holocaust survivors from 56 countries.

1 petabyte = 1,024 terabytes = 1,048,576 gigabytes
RE-IMAGINING CAMPUSES IN THE DIGITAL AGE

• Education: From teaching content to learning content AND competencies = Talent
• Research: From increasing specialization to specialization AND contextualization = the Research T
• Innovation: From Technology Transfer to Integrated Innovation (People-Centred)
“On a scale ranging from extremely interdisciplinary to exclusively disciplinary, how would you characterize your research?”

Source: 2008 web survey, SSH faculty, Science-Metrix
Canada’s Granting Councils (later 20th century)
NSERC

SSHRC

CFI

CIHR

21st century approach
Key observations about the emergence of Digital Scholarship

1. Students, professors, research partners and those in the larger society are now being connected formally and informally in efforts to learn about all aspects of the past and present, and to use such learning to help make a better future.

In this context, `research data` are now also `learning data` as well as `innovation data.`
2. Digital Scholarship is now evident across all fields of enquiry as data become the `coin of the 21st century realm.`

From studies of colliding particles to research on human thought and behaviour, the importance of Digital Scholarship is increasing rapidly across campus and beyond.
3. Data are now understood in terms of numbers, words, images, sounds, and, indeed, digital representations of all human and non-human phenomena.

In this context, the future will require increasingly sophisticated approaches/mechanisms for interrelating and integrating multiple datasets.
4. The use, re-use and re-purposing of data are becoming increasingly important articulations of the deep cultural changes now underway in education, research and innovation.
5. As distinctions between creators and users blur, data often exist in a dynamic rather than fixed state as a result of multiple and iterative engagement.
6. Complex privacy, confidentiality and ethical barriers arise around access to and use of personal and collective data across multiple researchers/research settings.

Since many legal and ethical considerations across diverse research areas are shared internationally, it is especially important that Canada engage with partners around the world in order to develop appropriate policies and practices for Digital Scholarship.
7. Most observers conclude that digital technology is far more advanced at the moment than either the availability of use-ready digital content or the support for users (including digital literacy education).

While continued improvement in computing processing and connectivity is essential especially given the increasing importance of massive data, special attention needs to be paid to enhancing access and use both on campus and beyond.

(One urgent need is for skilled and sophisticated people who can work effectively in a digital environment including technologies, content and literacies (from access to analytics)).
8. While discussion in the 1970s-1990s often stressed the importance of standardization in a 19th and 20th century cookie-cutter sense, the new emphasis is on both-and solutions to the problems of preservation, interoperability, meta-data, data delivery systems, user interfaces, etc.

In the distributed, open, empowered world of Digital Scholarship, coordination rather than control is key especially as networks replace vertically-integrated hierarchies.
9. Great unevenness characterizes the Digital Scholarship landscape especially with respect to preservation infrastructure which has become an urgent need to support and sustain digital scholarship.
10. In ways similar to the changes in the recorded music industry, digital scholarship is moving from an emphasis on data ownership to a provision of data services.

Whereas libraries once held and lent scholarly journals, for example, they now increasingly provide access to digital publications.

However, questions of ownership, curation, and access as well as business models for sustaining a `services` approach (that includes support for analytics, visualization, etc) remain far from resolved.
11. While the increased connectivity of digital scholarship has in many ways made physical location less relevant, the growing importance of massive data which cannot be moved easily has increased the importance of location.

In addition, the growing recognition of the importance of face-to-face interaction as a component of effective computer-enhanced learning and collaboration helps explain the importance of regional clusters with networked institutions that are then linked to remote locations nationally and internationally.
These key observations make clear the importance of updated and new policies and practices to coordinate the continued development of a robust and sustainable eco-system for enhanced Digital Scholarship.
HOW DO YOU BUILD A SUCCESSFUL DIGITAL MEDIA COMPANY IN CANADA?

YOU DO IT... WITH A NETWORK.

Entrepreneurs’ success enriches Okanagan tech scene

CDMN Canada 3.0 2013

Toronto, Ontario
May 14, 2013

Connect, Collaborate & Commercialize at CDMN Canada 3.0
Big Data Workshop

Building Effective Digital Infrastructure Towards Excellence in Research and Training

SSHRC CRSH

Organized by the Social Sciences and Humanities Research Council
Featuring:

Patrick Horgan
Vice President, Manufacturing, Development and Operations, IBM Canada

Ted Hewitt
Executive Vice-President, Social Sciences and Humanities Research Council (SSHRC)

Vincent Larivière
Assistant Professor of Information Science at the Université de Montréal

Dr. Wendy Cukier
Vice-President of Research and Innovation at Ryerson University and Founder/Director of Diversity Institute

Tom Jenkins
Executive Chairman and Chief Strategy Officer, Open Text Corporation; SSHRC Council Member
Neil Randall awarded $2.5 million from SSHRC to study games and player immersion

Monday, May 28, 2012 - 11:59am
uWaterloo's biggest SSHRC grant to date

What's in a game? If you ask English professor Neil Randall, he'll tell you: whole worlds are created in games, and games can be found all around us (even sometimes in contexts we don't expect). But in the end, it's the player's experience that matters most. “Understanding the player's experience, particularly in the increasingly immersive games made possible by digital advances and the Internet, is what interests me,” says Randall. “What are the factors (technical, psychological, sensory, social) that engage and motivate players and create a rich gaming experience? How might we use knowledge of these factors to create more sophisticated games, not only to entertain, but to educate, to train, to enhance social interaction, to simulate real-world scenarios, to improve health and wellness, etc.? And what are the specific factors or contexts associated with immersive game-play that can sometimes lead to problems, for instance addiction or social isolation?”
We cannot solve our problems with the same thinking we used when we created them.

-Albert Einstein
Large Hadron Collider
Over to you!

Thank You

À vous la parole!

Merci