KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

MEANING AND PERSPECTIVES IN THE DIGITAL HUMANITIES

A White Paper for the establishment of a Center for Humanities and Technology (CHAT)

SALLY WYATT (KNAW) AND DAVID MILLEN (IBM) (Eds.)

ROYAL NETHERLANDS ACADEMY OF ARTS AND SCIENCES
IBM RESEARCH
UNIVERSITY OF AMSTERDAM
VU UNIVERSITY AMSTERDAM
NETHERLANDS ESCIENCE CENTER

SALLY WYATT (KNAW) AND DAVID MILLEN (IBM) (Eds.)

Meaning and Perspectives in the Digital Humanities

A White Paper for the establishment of a Center for Humanities and Technology (CHAT)

CONTRIBUTING AUTHORS (IN ALPHABETICAL ORDER): Lora Aroyo, Rens Bod, Antal van den Bosch, Irene Greif, Antske Fokkens, Charles van den Heuvel, Inger Leemans, Susan Legêne, Mauro Martino, Nat Mills, Merry Morse, Michael Muller, Maarten de Rijke, Steven Ross, Piek Vossen, Chris Welty

Amsterdam, 2014

Royal Netherlands Academy of Arts and Sciences
PO Box 19121, NL-1000 GC Amsterdam
T +31 (0)20 551 0700
F +31 (0)20 620 4941
knaw@knaw.nl
www.knaw.nl

pdf available on www.knaw.nl

Typesetting: Ellen Bouma, Alkmaar

ISBN 978-90-6984-680-4

The paper for this publication complies with the ∞iso 9706 standard (1994) for permanent paper

FOREWORD KNAW

It is with great pleasure and pride that the Royal Netherlands Academy of Arts and Sciences (KNAW) and IBM Research, together with their primary partners University of Amsterdam and VU University Amsterdam, jointly present this *White Paper*. It combines a set of major challenges in the fields of digital humanities and cognitive computing into an ambitious public-private research agenda. Its goal, briefly put, is to develop a new generation of computer technology that is able to truly 'understand' products of the human mind, and past and present human activities. This certainly presents a challenge, as written text, still and moving images, speech or music – the core material of humanities research – naturally tends to be multifarious, ambiguous, and often multilayered in meaning; quite the opposite of the data that present-day computers prefer to crunch. We can only meet a challenge like this effectively by combining forces. This research agenda, therefore, has been drafted by a league of major players in this field, both humanities scholars and computer scientists, from IBM Research, the Royal Academy, and the two Amsterdam universities. Together, these four institutions have committed themselves to realize 'CHAT'– the Center for Humanities and Technology. This *White Paper* outlines the research mission of CHAT. CHAT intends to become a landmark of frontline research in Europa, a magnet for further public-private research partnerships, and a source of economic and societal benefits.

Prof. Hans Clevers
President of the Royal Netherlands Academy of Arts and Sciences

FOREWORD IBM

It is a great time to be reinventing the study of human behavior with innovative information technology. At IBM, we were pleased to be invited to collaborate with internationally recognized humanities and computer science scholars to create a research agenda for computational/digital humanities. Humanities data are both massive and diverse, and provide enormous analytical challenges for the humanities scholar. Deep, critical, interpretative understanding of human behavior is exactly the kind of problem that will shape the direction of the next era of computing, what we at IBM call cognitive computing. I look forward to seeing the results of a program of research, in which humanities scholars and computer scientists work side by side to understand important contemporary societal challenges.

Kevin Reardon
Vice-President, Corporate Development

TABLE OF CONTENTS

Foreword KNAW 5
Foreword IBM 6
Executive summary 8

Introduction 11

SECTION 1: OPPORTUNITIES FOR THE HUMANITIES 15

SECTION 2: CORE TECHNOLOGIES 19
Cognitive computing 19
Network analysis 22
Visualization and visual analytics 24
Text and social analytics 27
Search and data representation 29

SECTION 3: INFRASTRUCTURAL NEEDS 33
Architecture 33
Social infrastructure 38
Education and training 42

CONCLUSION 50

REFERENCES 51
APPENDIX: COMPUTATIONAL METHODS IN THE HUMANITIES 53

EXECUTIVE SUMMARY

The Royal Netherlands Academy of Arts and Sciences (KNAW), University of Amsterdam (UvA), VU University Amsterdam (VU), Netherlands eScience Center (NLeSC) and IBM are developing a long-term strategic partnership to be operationalized as the Center for Humanities and Technology (CHAT). The members and partners of CHAT will create new analytical methods, practices, data and instruments to enhance significantly the performance and impact of humanities, information science and computer science research. The anticipated benefits of this partnership include: 1) transformational progress in humanities research and understanding to address societal challenges, 2) significant improvements in algorithms and computational instruments that deal with heterogeneous, complex, and social data, and 3) societal benefits through novel understandings of language, culture, and history.

Four important humanities challenges have been identified, relating to the ways in which perspectives, context, structure and narration can be understood. More specifically, these can be expressed as: understanding changes in meaning and perspective, representing uncertainty in data sources and knowledge claims, understanding how patterns and categories are made and stabilized, capturing latent and implicit meaning, and moving from sentiment mining to emotional complexity.

We see many important scientific and technical challenges that promise breakthrough scholarship in the humanities. The work in CHAT will focus on the following areas: Cognitive Computing, Network Analytics, Visualization, Text and Social Analytics, and Search and Data Representation.

Cognitive Computing: Cognitive computing systems collaborate with humans on human terms, using conversational natural language as well as visual and touch interfaces. We envision cognitive systems will be able to learn and reason, to interact naturally with humans, and to discover and decide using deep domain knowledge.

Network Analytics: Contemporary network theory and technologies have transformative potential for the humanities by extending the scale and scope of existing work and by providing a framework for analysis.

Visualization: Access to large, multimodal datasets has created both challenges and opportunities for humanities researchers. New visualization instruments are required for interactive discovery of meaning across time and integrating multiple modalities. Progress is also needed in the underlying analytics on which these multi-layered visualizations are produced.

Text and Social Analytics: With current text analytics, we can perform computation of attributes of text, pattern detection, theme identification, information extraction, and association analysis. We envision ongoing improvements in lexical analysis of text, topic extraction and summarization, and natural language processing (NLP) of meaning and associations within the text.

Search and Data Representation: One of the key challenges in modern information retrieval is the shift from document retrieval to more meaningful units such as answers, entities, events, discussions, and perspectives. Advances in this area will help humanities scholars in important *exploration* and *contextualization* tasks.

In addition to the underlying core technologies described above, significant challenges are also present in both the computing and collaboration infrastructures to achieve the desired transformation of digital humanities scholarship. Hosted services for 'big data' must deliver easy access to both historical and 'born digital' data that comprise contemporary digital humanities research. Innovative collaborative platforms, social learning approaches, and new work practices are needed to promote and support data sharing and collaboration across multi-disciplinary, distributed research teams.

Transforming both humanities and computer science research will equip CHAT to contribute to meeting important societal challenges. Computers and computational methods, since their widespread development and diffusion in the second half of the 20th century, have transformed the ways in which people work, communicate, play, and even think. Humanities research contributes substantially to the development of the human spirit, and to critical reflection, especially in debates about social inclusion, multiculturalism, and the role of creativity in education, work, and elsewhere. Humanities research also makes

a vital contribution to many sectors of economic and cultural life, including media, heritage, education, and tourism. Humanities research has also contributed to innovations in computer technology, for example via predictive text, now available on all mobile devices. We believe that collaboration among humanities and computer science researchers in CHAT will lead to significant breakthroughs in both areas, and will benefit many other areas of social, technical and cultural activity.

INTRODUCTION

Recent advances in digital technologies have provided unprecedented opportunities for digital scholarship in the humanities. In this White Paper, we describe many of these advances and opportunities as background motivation for the development of an important multidisciplinary stream of research in the Center for Humanities and Technology (CHAT).

The promise of digital scholarship for the humanities has been articulated many times over the years. In her 'call to action' for humanities scholars, Christine Borgman (2009) argues that the transformation of the field will require new ways to create, manipulate, store, and share the many kinds and huge quantity of research data. Just as important, she adds that new publication practices, research methods and collaboration among researchers will be required. While much progress has been made in the field, more remains to be done. We need to go beyond improving access to data and knowledge, through digitization projects, in order to consider what kinds of new knowledge can be created using advanced analytic instruments and techniques (discussed in Section 2).

As part of an effort to invigorate and accelerate the transformation of humanities scholarship, several workshops have been held during the past two years in Cambridge, MA and Amsterdam. One of the goals was to reach a common understanding of some of the most important challenges within 'digital humanities' (DH).[1] Participants included researchers from the Royal Netherlands Academy of Arts and Sciences (KNAW), VU University Amsterdam, University of Amsterdam,

1 There are many terms in circulation: digital humanities, e-humanities, computational humanities, data-driven research, fourth paradigm, big data, etc. The choice often reflects subtle differences in emphasis which vary between linguistic and disciplinary communities, and over time. We take the view that all research and scholarship has already been changed by the widespread availability of digital tools for finding, collecting, processing, analyzing and representing data of all types (Wouters et al, 2013). We use the label 'digital humanities' only when we wish to make a particular distinction with humanities scholarship more generally.

the Netherlands eScience Center (NLeSC), and IBM.[2] Perhaps the most significant challenge identified was the need to acquire, represent, and archive humanities data in a way that is easily accessible to a broad range of scholars. Equally important is the need for powerful search and discovery instruments to enable scholars to explore humanities data from 'multiple perspectives'.

These workshops spurred wide-ranging conversations about specific projects and the instruments and research practices that were used. Much of the discussion centered on the kinds of humanities research currently underway using state-of-the-art instruments, and what innovations would be possible and desirable in this area. Often, the conversation focused on how new forms of humanities scholarship hold promise of new understanding of human behavior and potential for great societal influence and impact. We summarize the important humanities challenges and core technologies in Table 1, and describe each topic briefly below.

Table 1: Humanities Challenges and Core Technologies

Humanities Challenges	Core Technologies
Perspectives	Cognitive Computing
Context	Network Analytics
Structure	Visualization and Visual Analytics
Narration	Text Analysis/Topic Modeling
	Search and Data Representation

The humanities scholars described challenges in several important areas, including the following:

Perspectives: Current solutions for dealing with subjective information are limited to remote sensing of sentiment. Going beyond this requires deep language technology to automatically uncover and summarize different perspectives on current topics or cultural artifacts. The grand challenge is that the boundaries between perspectives are fuzzy and continuously shifting, both in mentions in text and realizations in images and video. Perspectives are also often in conflict, between individuals, groups and nations. Making these perspectives visible, and tracing their evolution could help in conflict situations, in diplomacy and in business.

2 Joris van Zundert made substantial contributions in the early meetings, for which we are very grateful. We would also like to thank the following for additional comments made during the preparation of this document: Patrick Aerts, Hans Bennis, Leen Breure, Peter Doorn, Christophe Guéret, Michel ter Hark, Theo Mulder, Andrea Scharnhorst, Frank van Vree, Demetrius Waarsenburg, and Henk Wals.

Context: The information landscape is changing. Instead of disconnected collections of individual snippets and snapshots, we now have access to longitudinal trails of utterances and ideas thanks to the digitization of our heritage and of our lives. The challenges here are: (1) how to determine such contexts (place, time, task, role) and how to determine the right granularity, and (2) how to chain such context to automatically identify complex activities ('buying a house', 'forming a project team', 'starting a business') and recommend answers and content based on this. Cognitive computers will be able to discover and present previously unknown connections between data and the importance of those connections.

Structure: The discovery of semantic and statistical structure is key to automatically assigning meaning and value to content. The ambition is to push the automatic creation of very large scale knowledge graphs that are populated with increasingly complex semantic units (entities, relations, activities, events).

Narration: Simple ranked lists of relevant documents leave it to humans to compose answers to complex questions. Our ambition is to create rich representations and build effective narrative structures, both textual and visual, from perspectives extracted from heterogeneous data.

The following important areas of technology that were discussed and considered important to enable new research in the humanities:

Cognitive Computing: Cognitive computing systems collaborate with humans on human terms, using conversational natural language as well as visual, touch and other affective interfaces. This partnership between human and machine serves to improve discovery and decision making by augmenting human abilities with brain-inspired technologies that can reason, that learn from vast amounts of information, that are tireless, and that never forget. We envision that cognitive systems will employ complex reasoning and interaction to extend human capability for achieving better outcomes. Cognitive systems will be able to learn and reason, to interact naturally with humans, and to discover and decide using deep domain knowledge.

Network Analytics: Contemporary network theory and technologies have transformative potential for the humanities by extending the scale and scope of existing work and by providing a framework for analysis.

Visualization: Access to large, multimodal datasets has created both challenges and opportunities for humanities researchers. New visualization instruments are required for interactive discovery of meaning across time and integrating multiple modalities. Progress is also needed in the underlying analytics on which these multi-layered visualizations are produced. Finally, new approaches are needed to communicate the stories that these visualizations reveal to audiences at all levels of visual literacy (i.e., consumable visualizations).

Text and Social Analytics: With current text analytics, we can perform computation of attributes of text, including determining word and n-gram frequencies, pattern detection, theme identification, information extraction, and association analysis, with a goal of turning unstructured text into data suitable for further analysis. Great progress has been made in linguistic and lexical analysis of text, topic extraction and summarization, and natural language processing (NLP) of meaning and associations within the text. Much remains to be done as these computation techniques are often fragile and incomplete, and require significant customization for each corpus. Nuanced language understanding (e.g. deception, humor, metaphor, meta-discourse) remains a significant challenge.

Search and Data Representation: One of the key challenges in modern information retrieval is the shift from document retrieval to more meaningful units such as answers, entities, events, discussions, and perspectives. Advances in this area will assist humanities scholars in important *exploration* and *contextualization* tasks.

This White Paper is structured into three main sections. Section 1 discusses current challenges and opportunities for *humanities* scholarship. Recent ground-breaking work in the area was considered and novel use cases were created to envision new directions. These form the background for Section 2 which describes the *core technologies* that are critical to the future of humanities scholarship. While much progress has been made in recent years in areas such as text analytics and cognitive computing, many technological challenges remain. Section 3 lays out some of the *infrastructural* challenges that lie ahead for breakthrough advances in this area, including the technology required, evolution of the social/collaborative infrastructure, and new forms of training and education for humanities scholars.

SECTION 1: OPPORTUNITIES FOR THE HUMANITIES

Developments in computational infrastructures, instruments and methods provide scholars in the humanities with many opportunities to collect, store, analyze and enrich their multimodal data (text, numbers, audiovisual, objects, maps, etc.), and to communicate their research data and results in exciting new ways. Such developments may also lead to new research questions, not only in the humanities but also in the computational disciplines. Furthermore, interdisciplinary dialog and cooperation have the potential to enrich the research programs of all involved.

CHAT will enable humanities scholars to both contribute to and take advantage of these developments, not only to address questions and challenges in their research fields and disciplines, but also to pioneer new forms of scholarship that bring together humanities and computational ways of thinking. It is thus important to keep a dual focus:

- Develop new computational instruments, methods, and approaches that can be used across a range of research questions and disciplines, and that address one or more of the challenges mentioned in the Introduction, namely perspectives, context, structure and narration.
- Understand how researchers can make effective use of such innovations in order to develop new research questions (see also 'social infrastructure' section), stimulate cooperation between academic, industry and other public partners, and meet societal challenges.

This dual focus will be brought together so that insights into how researchers actually use computational technologies inform the development of new instruments and methods, and to ensure that these are available to a broad group of researchers.

Scholarly Opportunities

Computational technologies offer scholars in the humanities many possibilities to find, manage, analyze, enrich, and represent data. The development of

instruments for, amongst others, text mining, pattern recognition, and visualization have potential benefits for the way in which humanities is conducted and for the questions researchers will be able to ask. During the preparatory meetings for this White Paper, participants were invited to prepare 'use cases', examples of research problems where new developments in computer techniques might offer some solutions, in relation to perspectives, context, structure and narration. These were the basis for extensive discussion (see Appendix for summary). Five important opportunities for the humanities emerged, each of which is briefly discussed below.

Understanding changes in meaning and perspective, over time and across groups

The ways in which humanities scholars understand historical and current objects will change as new sources come to the surface, and will depend on the scholar's own theoretical position and value system. Furthermore, understanding of the past is often largely informed by current issues and concerns. For example, understanding colonialism in different parts of the world has changed dramatically over the past 40 years. As new documents were discovered, its meaning was re-interpreted in light of discussions about empire, post-colonialism, and a new world order. History is full of such readings, which change over time and may differ among groups at any one time. It is not only 'grand' historical events that are subject to changes in interpretation. Single words, concepts, ideas and books can also have different meanings across time, space and social groups. For example, as a result of political action in the 1960s and '70s, 'gender' emerged in the humanities and social sciences as an analytic concept, opening up new areas of inquiry and requiring new interpretations of past events and documents. Another example is the ways in which canonical texts, such as the Bible or the teachings of Lao-Tzu, are subject to new readings by each generation of scholars. Developing instruments and techniques to help humanities scholars understand changes in meaning and perspective has obvious benefits in all areas of human communication.

Representing uncertainty, in data sources and knowledge claims

Changes in meaning and perspective arise from the availability of sources and reference material. Humanities scholars have traditionally had access to books, documents, manuscripts, and artifacts held in libraries, museums, and archives. A fundamental part of the training of humanities scholars is to learn to question the provenance and representativeness of the sources available (Ockeloen et al, 2013), and to ask questions about what might be missing, whose voices

and opinions are included, and whose might be left out. As the data and sources become increasingly digitized, it is important to develop new ways of understanding and representing the nature of the data now available and the claims being made. Again, this is a pressing issue for humanities scholars, but is of wider relevance, especially as techniques for visualization become ever more sophisticated, and as the available data varies so much in quality.

From patterns to categories and interpretation and back again
Some humanities scholarship is based on identifying and explaining the exceptional person, object, or event, often as a way of opening up bigger questions. Scholarship is also concerned with the search for patterns, trends, and regularities in data about historical occupations and disease, or the use of particular words in a novel or poem. Identifying such patterns can result in the development of categories for further analysis and use, but these categories may then become too rigid, leading later researchers to miss important new patterns or novel exceptions and outliers. (Bowker and Star, 1999) Developing instruments to allow for multiple categorizations, as new data become available, is important not only for humanities scholars but for all who deal with big datasets. This is especially challenging for historical data sources, where the data is often incomplete and heterogeneous. The ways in which data can be combined and recombined to make categories is important not only for researchers but also for policy makers in all fields, who are seeking meaningful classifications of occupation, crime, and disease.

Capturing latent and implicit meaning in text and images
While the availability of databases with pre-existing named entities is a valuable resource for many types of research, sometimes scholars aim to understand more latent and implicit dimensions and meanings of text and data, such as irony, metaphors and motifs. This fits well with current developments in topic modeling that is language-independent, and which is based on stochastic modeling and information theory (Karsdorp and van den Bosch, 2013). Such developments would have applicability in a range of sectors, including courts, marketing, and anywhere where nuance in meaning is fundamental to interpretation and action.

From sentiment mining to emotional complexity
Huge advances have been made in recent years in 'sentiment mining' of contemporary digital material. However, this characterizes utterances as positive, negative, or neutral. Yet human emotions are much more complex, and are

expressed not only in words but in gestures, expressions, and movements. In addition, linguistic and body language changes across time, gender, ethnicity, nationality, religion, etc., such that it makes sense to talk of 'emotional communities', each with specific styles and practices. Humanities sources, including literature and artistic works, provide a rich resource for developing a fuller and more nuanced set of emotional classifications.

Organizational opportunities

Humanities scholars have a long history of engagement with computational technologies (Bod, 2013). Yet adoption of advanced analytical instruments and methods remains limited. Tools are sometimes developed but, due to lack of long-term funding, are not maintained and thus quickly become out-of-date. Not all sources are available digitally, and there remain barriers facing those scholars who work with material that has not yet been digitized. For a variety of reasons, many scholars in the humanities remain unaware of the potential benefits of applying computational methods to their research.(Bulger et al, 2011) CHAT can address these, via the following mechanisms:

- implement the lessons of previous work in the area by the contributing partners
- improve understanding of the needs of humanities researchers
- improve awareness of the potential and availability of computational instruments and methods
- promote policies for the preservation of computational instruments and data for future researchers and for the digitization of currently analog research material
- engage with a range of potential partners in the cultural heritage sector and creative industries
- contribute to debates about the future of humanities and the role of computational technologies

CHAT will work toward achieving *complex*, *computer-based* and *cooperative* humanities[3]: complex in terms of questions, data, and interpretations; computer-based data, methods and representations; and, cooperative across disciplines, sectors and countries.

3 This is inspired by the recent issue of *De Groene Amsterdammer* (31 oktober 2013) which looked at the 'ten revolutions in the humanities'. Digital humanities figured prominently, as did alliteration.

SECTION 2: CORE TECHNOLOGIES

To address the opportunities for the humanities outlined in the preceding section, five core technologies have been identified for development within CHAT: Cognitive Computing, Network Analysis, Visualization and Visual Analytics, Text and Social Analytics, and, Search and Data Representation. Each of these technologies is described in terms of the current state-of-the-art as well as challenges and future possibilities.

COGNITIVE COMPUTING

A Cognitive System is a collection of components with the capability to learn and reason intelligently. By this definition, people are cognitive systems, but so are teams, companies, and nations. Until recently, computers were not cognitive systems. They were instruments employed by human and group cognitive systems to enhance their own function, and were expected to be perfectly correct in their operation. The success of IBM's Watson system at beating the best human champions at the game show *Jeopardy!* represents a new level of cognitive computing achievement. Unlike past successes such as the famous Deep Blue chess match with Gary Kasparov, the *Jeopardy!* challenge engaged the system in a contest involving human language with all its variation, imprecision, and ambiguity. This required vast amounts of knowledge, reasoning, and the ability to evaluate its confidence in its conclusions. The emergence of cognitive computing will enable computers to be better instruments, but also to function as collaborators and participants in and contributors to larger collective intelligence cognitive systems.

Cognitive computing is in part a reaction to the growing data deluge that affects all sectors of modern life. Just as the rate of information flow long ago increased beyond the limits of individual humans to keep up, it is now outstripping the ability of computers to handle. Despite their still increasing capacities

(even if the rate of increase is slowing as fundamental physical limits are reached), current computers cannot keep pace with the ever-increasing flow of data. The von Neumann processor-centric architecture, which has served us well for over 60 years, must ultimately be replaced by a more distributed data-centric architecture in which data and computation are distributed throughout the system. Watson's highly parallel architecture is a step in that direction. Even more extreme solutions are under development, such as the Synapse chips that will enable the creation of cognitive systems out of collections of artificial neurons.

Cognitive systems such as Watson have moved beyond the approach of providing a single algorithmic approach toward solving a problem. Instead, they deploy many different approaches, often in parallel, using a variety of information sources for generating hypotheses, and then employ a variety of different approaches and knowledge sources to score, rank, and choose among them. This hybrid approach does not guarantee correct answers, but as Watson proved on *Jeopardy!*, can achieve results exceeding human performance.

Current technology

The state of the art represented by Watson has moved beyond the familiar keyword-based search engine functionality. Watson is ultimately a question-answering machine. Rather than simply searching for potentially relevant sources of information based on the words in the question, Watson attempts to understand the question, determine what sort of answer is called for, find candidate answers in the vast trove of information that it has ingested, and finally evaluate the confidence it has in its potential answers. Since the *Jeopardy!* win in 2011, progress has continued. In the two and a half years since Watson outscored humans on *Jeopardy!,* work has been proceeding to apply cognitive computing capabilities to other domains, such as medicine and finance. While the original version of Watson could display multiple answers under consideration along with their confidence scores, the new Watson can provide justifications for its answers, providing access to the evidence that supports its conclusions. The ultimate goal is for cognitive computer systems to be able to engage in natural language conversations with people, allowing a give and take exploration of a problem or topic, dealing with the issues of uncertainty raised in Section 1.

Challenges

Existing cognitive computing systems are large machines with voracious energy requirements. The original Watson system consisted of a cluster of 90 IBM Power750 servers with a total of 2880 3.5 GHZ POWER7 processor cores and

16 Terabytes of RAM, enabling it to store 200 million pages of structured and unstructured information. It consumed 85,000 watts of electricity when running at full speed. As technology improves, the size and energy requirements of such systems will likely decrease substantially. At present, training the system, locating, obtaining, curating, and feeding in the millions of pages of data from which it will draw its conclusions, is a time-consuming and effort-intensive task.

More importantly, the most significant challenge in advancing cognitive systems like Watson is adapting the technology to new domains. While the *Jeopardy!* challenge was in many ways "domain independent", the knowledge required to answer these questions was fairly shallow. For answering more domain-specific questions, such as those in finance, law, and medicine, as well as for all humanities disciplines, understanding the specialized vocabulary and dealing with the lack of a measurable ground truth, is a key challenge. Humanities questions, both general and specific, rarely have a single, correct, answer. Rather, a range of resources may be relevant to help the human-machine collaboration find the desired results.

Possibilities

The question of how cognitive systems could be applied to the humanities is necessarily a speculative one. Unlike other more established areas, such as text analytics and visualization, cognitive computing is so new that there are no existing humanities applications from which to extrapolate. Nevertheless, there is considerable potential. The capacity to ingest millions of documents and answer questions based on their contents certainly suggests useful capabilities for the humanities scholar. The ability to generate multiple hypotheses and gather evidence for and against each one is suggestive of an ability to discover and summarize multiple perspectives on a topic and to classify and cluster documents based on their perspectives.

While we do not expect a cognitive computing system to be writing an essay on the sources of the ideas in the *Declaration of Independence* any time soon, it might not be unreasonable to imagine a conversation between a scholar and a future descendant of Watson in which the scholar inquires about the origins and history of notions such as 'self-evident truths' or 'unalienable God-given rights'[4] and carries on a conversation with the system to refine and explore

4 The famous sentence from the US Declaration of Independence, adopted on 4 July 1776, drafted by Thomas Jefferson, is: *We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.*

these concepts in the historical literature and sources. The resulting evidence gathering and the conclusions reached would be a synthesis of human and machine intelligence.

Application of cognitive computing capabilities beyond written or spoken language interactions, such as analysis of images, video, or sound, is another potentially fertile area of research that could provide great value to the humanities. Cognitive computing evaluation of images of paintings or audio recordings of musical performances could provide insight into the identity of unknown painters or composers, or assist in analysis of artistic influences or changes in technique occurring over time.

Consideration of the concept of aggregate cognitive systems suggests other possible applications in the humanities. We can go beyond simply providing computational infrastructure support for collaboration and communities of interest, and enhance the group cognition processes of collections of human and computer science scholars to address difficult humanities-related problems.

NETWORK ANALYSIS

Current technology

Like many fields, humanities have long been concerned with relationships that can be summarized and analyzed as networks. Most obviously, social networks connect people with one another. Many humanities scholars also study document networks, such as linked letters. Trade routes may be thought of as another type of network, with both geographic and economic relationships. Linguists create network representations of words and meanings, and study the relationships between language variants across geographies, human migrations, ecological changes, and colonialism. Historically significant epidemics also traveled over networks of various kinds – geographical, trade, imperial, and so on.

Contemporary network theory and technologies have transformative potential for the humanities by extending the scale and scope of existing work and by providing a framework for analysis. These approaches have already been brought into some humanities research programs. Literatures of national origins present gods, heroes, and (often) monsters who interact with one another, and who follow historical or legendary paths and types of relationships. Characters in complex narratives, such as Shakespeare or the conquests of Alexander the Great, engage in network relationships. Analytic approaches to networks permit computation of the relative positions of actors in a network,

as well as the strengths of their relationships. Through different types of network metrics, we can discern who is a crucial intermediary, or who is central to a conspiracy.

Networks are, in general, composed of nodes and links between nodes. In a structural network analysis, the nodes are treated as relatively interchangeable – we do not care about their attributes. Social network analysis enriches the structure with details about each node in the network, such as her/his nationality, class, sexual identity, or political party. Nodes may also be things, such as documents, in which case a document may have attributes such as authorship, readership, and genre. Hybrid networks are composed of two or more types of nodes. Humanities scholars may be particularly interested in hybrid networks of writers, readers, publishers, and documents, or of painters, paintings, patrons, and art dealers. In addition, the links between networks may have direction (e.g. an author writes a document, and a reader is influenced by the author through the document) and magnitude or tie-strength (e.g. lovers may have stronger ties than acquaintances). Formal network analysis unites the individual cases of persons, objects, and their relationships. In return, lessons learned from the individual cases may suggest new hypotheses and can sometimes be used to restructure for formal network analyses.

Challenges

While there are many tools available to visualize a network, the tools to construct the base data for such a network visualization remain complex. Broad usage is limited by the need to understand some mathematical formalisms – this is especially true of the network analytics (e.g. the different forms of centrality). Definitions of network concepts, such as 'brokerage' in 'betweenness centrality,' often carry unintended cultural assumptions (e.g. the Western valorization of individuality and uniqueness for the 'gatekeeper' options of a 'broker' between otherwise unconnected people).

From a disciplinary perspective, there is insufficient support in typical network analytics and visualizations for the tendency among humanities scholars to examine, compare, and combine multiple perspectives and interpretations. As Drucker (2011) has commented, many network formalisms are concerned with certainty and authority, not the layering of contingent and contextual meanings that are a focal concern of the humanities (see also section on 'Social Infrastructure'). Innovative work is needed to reshape the existing networking technologies and concepts to support more questions and analyses in the humanities.

Possibilities

In the future, we envision a suite of network analytic instruments, suitable for humanities data and questions. These instruments will be broadly available and interoperable. They will support derivation of network structures from the kinds of mark-ups that are already a part of digital humanities practices, such as enhanced versions of TEI, OAC, and YAML (see section about Text and Social Analytics). The existing network analytics, which have already proven useful in representing history and commerce, will be extended to support specific in the humanities needs emerging from more formal analyses of narratives and poetry. Concept networks will become more important in the analysis of genres, argumentation, and close readings of literary works. The humanities can increase the scope and scale of their work and their impact, and can inform network thinking by bringing specifically humanities-based concepts into the broader discussions of network analysis and representation.

Relations in networks are more and more interpreted and typed, eventually leading to formally structured graphs in representations such as RDF, VNA, or DL. These representations lead to an Open Linked Humanities infrastructure, in analogy to the Linked Open Data project (LOD), in which data are semantically anchored and linked. Such an Open Linked Humanities framework allows for news ways of network analysis (e.g. exploiting semantic generalization) and visualizations, such as graph exploration, timelines, and interactive maps.

VISUALIZATION AND VISUAL ANALYTICS

Within the humanities, data often come from sources different to those used to solve business and scientific problems. They may derive from metadata capturing the attributes of a collection of archeological artifacts, or text analytics run over a corpus of documents, or from any number of other humanities research sources, but these data are of themselves of little value unless they ultimately contribute to understanding. Visualization and visual analytics are means of extracting understanding from data.

Visualization is part of a larger research process that often begins with a question. The researcher will then need to determine what data, analytics, and visualizations to use to attempt to answer that question.

The visualization part of this process transforms data, which could come from databases or from analytics running over databases, sensor streams, unstructured textual resources, or structured metadata repositories, into some sort of graphical representation. Presenting data in a visual form allows us to

take advantage of human visual processing capabilities that have evolved over millions of years to enable us to detect trends, patterns, and configurations in the world around us. Although visualizations can be used merely as illustrations, the real power of visualization is its ability to make powerful arguments, provide insight, and raise new questions.

Visual Analytics is the application of interactive information visualization technology combined with computational data analysis to support the reasoning and sense-making process in order to draw better and faster conclusions from a dataset.

Current technology

Techniques for visualizing data have been under development for eons, dating back to Stone Age cave paintings that displayed information about animal populations and constellations (Friendly, 2006). Making maps that capture geographic information is a practice that goes back thousands of years. Timelines, line graphs, bar charts, and pie charts were introduced in the 18th century. A variety of more recent innovations have expanded the options available for visualizing data, including word clouds to summarize the frequency of common terms or themes in documents or corpora, network diagrams to represent entities and relationships between them, tree maps to visualize hierarchically structured data, and several means of visualizing high dimensional data. The advent of powerful computers with graphical capabilities makes possible nearly instantaneous generation of graphs and charts, animation, and interactive exploration of data via manipulation of the visual representation.

A variety of tools is available for the visualization of data, such as IBM's ManyEyes and MIT's Simile, open source tools such as GGobi and D3, and proprietary tools such as Adobe Flash and Tableau. Tools such as these have been applied to a wide variety of problem domains, including many relevant to the humanities. Projects like Stanford's *Republic of Letters* (Chang et al, 2009) allow interactive exploration of data from the 'Electronic Enlightenment' dataset around the exchange of correspondence in 16th through 18th century Europe demonstrate how massive tables of data can be brought to life and made amenable to exploration through visualization. A series of coordinated multi-view visualizations of a metadata repository with spatial, temporal, and nominal attributes, allow scholars to explore different aspects of this rich dataset, compare the correspondence networks of different authors, and view animations of the flow of information through this historical social network.

Challenges

Every new area that can be explored raises questions about what data or aspects of the data to visualize, what sorts of data transformation will be useful, and what metaphors or types of visualizations to apply. The proper answers to these questions depend on the user, the perspective, the audience, the context, goals, nature of the data, and the device or devices to be employed. Making these kinds of decisions is a skill that must be learned. It is easy to draw misleading conclusions from improperly applied visualizations.

Increasingly large datasets also present a challenge. While small datasets can be easily processed on a scholar's local workstation, very large datasets cannot be quickly transmitted and processed locally, making it difficult to support interactive exploration at high resolution.

Visualizing temporal aspects presents unique challenges. Capturing and making visible changes over time at speeds that users can consume and understand is still more art than science.

While there are a number of existing visualizations for textual data, there is still much room for innovation in this area. There is much promise in combining advanced text analytics and sophisticated visualization technologies. One can imagine how efforts such as the *Republic of Letters* project could be further enhanced with the addition of the right content-based visualizations. Indeed, projects such as CKCC (Roorda et al, 2010) have made a start in this direction, but there is much more than can be done.

Often the data we collect are imprecise, subject to error, or collected in a way that highlights certain aspects and makes invisible or de-emphasizes other aspects of a particular topic. Predictive analytics introduce additional uncertainty. Effective presentation and communication of data or analytic results in a way that comprehensibly conveys the inherent uncertainty continues to be a thorny issue.

Possibilities

New types of visualizations are being developed regularly, often spurred by the requirements of new datasets or needs. Ideally, these special purpose visualizations can later be generalized and applied to other domains.

The current state-of-the-art in visualization and visual analytics requires familiarity with available instruments and a deep understanding of the structure and characteristics of the data in order to make progress. Often this requires a close collaboration between computer scientists and humanities scholars to build instruments to enable exploration of specific datasets, as was the case with the *Republic of Letters* project. We look forward to a data repository with

metadata describing the data structure and contents integrated with a library of semantically described analytic methods and a cognitive computing infrastructure capable of reasoning with these descriptions and interacting with the user. This could make possible an interactive discourse where appropriate datasets were selected, transformations applied, and visualizations chosen in a cooperative collaboration between scholar and computer system. Large displays and immersive environments, along with conversational speech, touch, and gesture recognition could make creation and exploration of visualizations easier and more natural.

Recognizing that scholarship is increasingly a collaborative venture, we can see the benefit of integrating interactive visualizations within a collaboration infrastructure so that the visual analytics being applied could support collective as well as individual reasoning. This would allow scholars to share visualizations as live views of the actual data, to provide evidence or raise questions, allow their collaborators to explore those data and visualizations from other perspectives, and also share the insights that they discover.

TEXT AND SOCIAL ANALYTICS

The exponential growth of textual resources over the past 600 years has made it impossible for scholars to read all the material available on almost any topic, no matter how narrow. At the same time, the explosion of computer power available to the individual researcher, as well as the trend toward digitization of textual materials and the development of a world-wide digital communications network, has opened up new ways to analyze written works, and created opportunities to study large text corpora.

Text analytics perform computations on attributes of text, such as determining word and n-gram frequencies, performing pattern detection, information extraction, and association analysis, with a goal of turning unstructured text into data suitable for further analysis. While the application of text analytics is no substitute for 'close reading' of a text, it enables a kind of 'distant reading' survey of large amounts of text for purposes of establishing an overview and detecting large scale or historical patterns, and for pinpointing particular works or sections of works among a large corpus to be subjected to further close reading.

Text analytic techniques run a spectrum from application of purely statistical methods such as tracking word frequencies in documents to more advanced natural language processing techniques including stemming, part of speech tagging, syntactic parsing, and other deep linguistic analysis approaches to

achieve named entity recognition, event detection, co-reference identification, sentiment analysis, and underlying semantics.

Current technology

The state-of-the-art in text analytics today includes a variety of capabilities that may be of use to the humanities scholar. Text retrieval functionality, familiar to billions of World Wide Web users, makes it possible to find texts, or sections of texts based on particular search words or phrases. Instruments are available that compute term frequencies, counts of words or phrases in documents or sections of documents. These can be used to characterize text passages, detect themes, and give a crude overview of a section, document, or corpus. A variety of statistical algorithms can distill this information into topic models that characterize the text in terms of a small number of theory-informed distinctive words or phrases, analyze word frequencies and use patterns to establish authorship or attributes of the authors, or analyze the evolution of language use over time. Summarization algorithms can attempt to extract or synthesize key sentences that convey what a particular document, passage, or corpus is all about. Metadata about textual works, such as author, date, and location, support analyses such as mapping the spread of ideas over time and space, or tracking influence across a social network of authors, editors, and readers.

Different combinations of these text analytic technologies have already made possible many interesting accomplishments that begin to illustrate the kinds of capabilities that are now available to the humanities scholar. The authorship of the twelve disputed Federalist Papers, for example, was established through analysis of word frequencies in the papers and comparison with the Federalist Papers whose authorship was already known. (Mosteller and Wallace, 1964) A variety of analyses pinpointed James Madison as the likely author. Combining text analytic results with data visualization technology is often a fruitful way to make sense of what the analytics are telling us. In this way, it was possible to map the spread of accusations that occurred during the Salem Witch Trials, and see the spread of mass delusion in a manner quite akin to the spread of disease (Ray, 2002). In another example, topic modeling was used to track the rise and fall of themes in Benjamin Franklin's *Pennsylvania* Gazette from 1728 to 1800 (Newman and Block, 2006). The Text Encoding Initiative (TEI) consortium has spent the past ten years developing and maintaining a standard for encoding machine-readable texts in the humanities and social sciences that is capable of capturing many forms of valuable metadata. The Online Archive of California (OAC) has developed a different set of mark-up standards. YAML offers a less-constrained mark-up standard.

Challenges

Despite the tremendous progress and capabilities available today, there is still much room for expansion and innovation. Existing repositories of textual material are scattered, unlinked and incompatible. Competing mark-up standards may exacerbate the incompatibilities. Most text analysis tools in use today use very shallow techniques that make no attempt to understand the content being analyzed, and are not dealing with syntax, inflected forms, categories of objects, synonyms, or deeper meaning. There is limited machine-readable metadata associated with most text archives. While primary sources are often digitally curated, scholarly works about those sources are seldom digitally available, and especially not in a form that is linked or linkable to those sources.

Possibilities

We can envision the future emergence of large and linked interoperable text repositories that have vast amounts of text available for analysis, as well as standards for capturing annotations, insights, cross references, hypotheses, and arguments in a sophisticated linked metadata tied to and accessible through digital texts. Deep analytics that make use of linguistic and semantic knowledge will allow more thorough and insightful analyses, and conversational interaction with cognitive systems with access to all this data and metadata will enable scholars to interactively describe, refine, and conduct a wide range of analyses on everything from sentences to large corpora.

Text analytics technologies do not in any way replace the work of the humanities scholar. Instead they provide a set of new instruments that can be wielded by the scholar to undertake analyses and achieve insights that would not have been possible otherwise.

Search and Data Representation

Current technology

Information retrieval, the scientific discipline underlying modern search engine technology, addresses computational methods for analyzing, understanding and enabling effective human interaction with information sources. Today's web search engines have become the most visible instantiation of information retrieval theories, models, and algorithms. The field is organized around three main areas: (1) analysis of information sources and user behavior; (2) synthesis of the heterogeneous outcomes of such analyses so as to arrive at high quality retrieval results; and (3) evaluation, aimed at assessing the quality of retrieval results.

In terms of analysis, considerable attention has been devoted to different ranking models based on the content of the documents being ranked, on their document or link structure, or on semantic information associated with them, either as manually curated metadata or derived from linked open data sources. Explicit or implicit signals from users interacting with information are increasingly being used to infer ranking criteria.

Multiple ranking criteria are being brought together to arrive at an overall ranking of documents. In recent years, there has been a steady shift from supervised mixture and fusion methods to learning, including using manually labeled data, and rank-based methods using supervised approaches. Semi-supervised or even unsupervised methods are beginning to emerge.

Finally, there has been a gradual broadening of the available repertoire of evaluation methods. The field of information retrieval has a long tradition in offline evaluation, where labeled datasets, created using expert or crowd sourced labels, are used to assess the quality of retrieved items, often in terms of metrics based on precision and recall. This tradition has been complemented with user-centered studies, in which users of an information retrieval system are observed in a controlled lab environment. In recent years a third line has been added: online evaluation in which experiments are being run, and implicit feedback is being gathered, with live systems, using methods such as A/B testing and interleaving.

Challenges

One of the key challenges in modern information retrieval is the shift in focus from document retrieval to information retrieval: in other words, the unit is shifting to meaningful units such as answers, entities, events, discussions, perspectives. On top of that we are seeing an important broadening in the type of content to which access is being sought: not just facts or reports of factual information, but increasingly also reports of opinions, experiences, and perspectives on these. This creates numerous search challenges.

For instance, in content analysis, new ranking algorithms are being sought that are able to differentiate between various implicit ways of framing a story. And since sources containing social and cultural information are highly dynamic, semantic analysis is challenging as open knowledge sources may be incomplete and not yet cover the entities being discussed in social media (Bron, Huurnink, and De Rijke, 2011). Understanding the way people describe images and videos (Gligorov, et al, 2011) and the influence different types of annotations have on precision and recall in search results (Gligorov, et al, 2013), is yet another challenge for search and recommendation systems. Another challenge

is that, today, we understand very little about the ways in which people, either as consumers looking for entertainment, or as professional researchers pursuing academic goals, search and explore social media data.

In terms of synthesis, there is a clear need to develop online ranking efficiently, with methods that need relatively few interactions and thereby open up the way to personalized ranker combinations that take into account the background and knowledge of individual researchers. We also need to understand how such methods best aggregate structured and unstructured retrieval results (Bron, Balog, and De Rijke, 2013).

New metrics for search results are being sought to fit the openness and diversity of web collections (where recall often does not apply) and to align directly with people's drive to discover new data and combinations of data by chance and serendipity (Maccatrozzo, 2012). Additionally, online crowds are being employed to evaluate large amounts of search results, as well as providing gold standard data covering a variety of features for systems to train. As the very nature of human information consumption is based on subjective interpretations and opinions, it poses new challenges for the evaluation and training process when it is recognized that there is not only not a single correct answer, but also that the correct answers are not always known (Aroyo and Welty, 2013).

Possibilities

Two main activities that humanities researchers engage in when using large collections of digital records in their research are *exploration* (developing insight into which materials to consider for study) and *contextualization* (obtaining a holistic view of items or collections selected for analysis). In the humanities, the approach to research is interpretative in nature. To shape their questions, researchers embed themselves in material and allow themselves to be guided through their knowledge, intuitions, and interests (Bron et al, 2012). Finding patterns in structured background knowledge, such as linked data sources, opens possibilities to study the diversity of contexts and their corresponding influence on measures for relevance and ranking, for example in recommendation systems (Wang et al, 2010).

The term *contextualization* refers to the discovery of additional information that completes the knowledge necessary to interpret the material being studied. For example, when studying changes in society over time by examining news broadcasts, the dominance of reports about crime would suggest that society is unsafe and degenerate. Understanding the news production environment, however, provides an explanation in that crime is covered constantly because of its popular appeal (Bron et al, 2013).

How do we best support multiple perspectives in the exploration and contextualization activities of humanities scholars? Innovative ranking methods, based on criteria to be elicited, implicitly or explicitly, from humanities scholars, are an important ingredient. We need innovative presentations of results organized around semantically meaningful units, such as answers to questions or cultural artifacts.[5] Likewise, there are opportunities for result summaries that describe not just the content but also more subjective aspects (positive vs. negative, or subjective vs. objective). And finally, we need insights on (1) the information preferences of users in their media choices and information consumption, and (2) operationalization of new concepts such as diversity, serendipity, and interpretative flexibility that could be used for information filtering, clustering and presentation in specific contexts.

5 KnowEscape is a COST Action (2013-17), involving the KNAW and other European partners. It brings together information professionals, sociologists, physicists, humanities scholars and computer scientists to collaborate on problems of data mining and data curation in collections. The main objective is to advance the analysis of large knowledge spaces and systems that organize and order them. KnowEscape aims to create interactive knowledge maps. End users could include scientists working between disciplines and seeking mutual understanding; science policy makers designing funding frameworks; cultural heritage institutions aiming at better access to their collections. (http://knowescape.org)

SECTION 3: INFRASTRUCTURAL NEEDS

In order to realize the potential of combining humanities and computer research, a number of infrastructural conditions need to be met. In this section, we discuss three: technical architecture, including repositories for data and instruments, and interoperability between datasets; social infrastructure to support collaboration across time, distance and discipline; and, education and learning for current and new generations of researchers.

ARCHITECTURE

Support for research requires distributed electronic access to a vast virtually centralized repository containing a variety of humanities artifacts and information about them. By simplifying access to the original artifacts as well as promoting contributions of insight about them, such a repository should promote humanities contributions to understanding past, present, and future human culture and behavior.

Creating, maintaining, and managing such a repository presents challenges in a variety of domains. Information has diverse formats, which may require massive storage. Some data are unstructured, requiring restructuring to become searchable. Portions of the repository may require restricted access. Content is open to many interpretations (some contradictory), and deliberation of these interpretations may prove as insightful as the original artifact. The success of the site depends on ease of access and acceptable response time, and such success could generate more traffic that may stress these success factors. Finally, keeping the content secure and up to date requires ongoing attention. We will explore each of these challenges and their impact on architectural choices needed to design a well performing knowledge repository.

Data diversity

Historic humanities artifacts are physical objects, so the choice for how they should be portrayed as electronic media may affect how they may be used for research. Three-dimensional artifacts may require exploration using an interactive 360° panoramic viewer, however there may be needs for additional, non-visual metadata attributes like weight, composition, and provenance. Each artifact may require multiple representations, each carrying their own, possibly controversial, interpretation For example, translation to a modern language may render an artifact easier to understand, but the translation may be evolving (Venuti, 2013). Therefore, each artifact will likely require extensive metadata to describe its attributes and facilitate search, visual representation to provide inspection, and audio describing the history and significance of the artifact (perhaps synchronized with video). Accompanying each artifact's representation would be deliberation and rationale developed to make claims about the artifact or its representation.

Web technology standards continue to improve and be refined, so there are considerations for migrating formats of artifact representations to ensure that they can be consumed. Though most web advances have not made prior formats obsolete (though new formats may perform better), this is not true for storage media. In cases where large amounts of data have been accumulated, and transfer across the Internet is not practical, the choice of medium and compression techniques to reduce content size may change over time. Likewise, if such data cannot be stored on disk due to size and/or frequency of access, the backup/archive medium must be upgraded from time to time to ensure it remains viable and accessible.

Today, data volumes are exploding due to the variety of sensors and intelligent devices. The repository should be designed to accommodate current cultural artifacts in addition to historic ones. This may open the repository to become a federation of sites that share and cross reference content. This adds complexity depending on whether content is to be sent directly from the hosting site or redirected through the point where the consumer has access to the virtual repository. Support and availability service levels would need to account for multiple sites, organizations, and communications between them.

Finally, such a repository will be more valuable if it continually adapts to new insights and can accommodate input from its community (see 'Social Infrastructure' below). Alternative representations, analytic results, papers, presentations, and talks related to artifacts should be able to be added to the repository content (with proper vetting and integration with metadata). This

may add requirements for data cleansing, transformation to current repository storage standards, and positioning relative to existing repository artifacts (e.g. is this a replacement for a prior representation). It may mean the repository will need to preserve the provenance of its content and store its content in a temporal, versioned repository (to allow access to or restoral from prior versions).

Data organization and storage

Many historic artifacts are initially represented as unstructured data, including but not limited to free form text, audio, video, or painted, written, or woven content preserved as images. To some degree, analytics on the content may be used to mine metadata necessary to identify or classify the content. Additionally, people will manually add classifiers to help support search and organization. While there are traditional ways to organize data using relational data stores for well-defined data structures, there are additional choices to hold semi-structured content to accommodate evolving choices for its classification.

NoSQL based data repositories allow self-described, text-based content to be stored and automatically indexed based on their metadata / schemas. This includes newer archive formats like JSON. We envision the need for both SQL and NoSQL databases to store content so it can be made available as quickly as practical without delay for the search attributes to be developed. As new metadata descriptors are introduced, they may be retroactively applied to existing artifacts.

Semantic metadata can be stored in an RDF repository, using a linked-data strategy that will allow us to link to and augment data from other sites, and support complex metadata search. This will enable data from different sites and sources to be queried, and ultimately for reasoning over semantic metadata to take place. Using RDF and/or REST services can allow open access to data we manage, allowing us to share as well as consume content from others.

We anticipate that a semantic taxonomy for the metadata will greatly simplify search by allowing gross level to fine level resolution searches. This implies a need for managing the taxonomy and promoting its reuse to avoid accumulating myriad synonyms for search terms. Assuming that the repository will support multiple languages, the taxonomy should provide for mirrored concepts in all languages. The growth and maintenance of the taxonomy should allow authors submitting new content to choose from existing terms as well as to introduce new terminology. The new terms should go through a vetting process to keep the taxonomy focused and to minimize redundancy.

Transformation and analysis of data should be automated as much as possible to speed access to artifacts, and for consistency of artifact classification. In cases where there are vast quantities of data to be analyzed, the speed of analysis can be improved if the analytics support is run in parallel using Hadoop-based repositories. As new analytics are introduced they should be run against the existing content to upgrade the knowledge about the artifacts.

Paying attention to how people specify their search needs and accept or reject results may provide insights about the data organization scheme. Providing dialoguing facilities to assist with metadata selection from the taxonomy and to refine facets used to filter results can assist people to find what they need.

Data access

Assuming considerable storage requirements to support the repository and the need for high availability, the serving infrastructure needs to perform well while providing flexibility and resiliency in case of failure. Storage can be organized in a SAN though this tends to help with latency and not bandwidth. Caching and POSIX compliant General Parallel File Systems (GPFS) can provide superior alternatives to the Hadoop Distributed File System (HDFS). The goal is to organize data with redundant storage to speed access and limit cross network / bus reads while providing resiliency should one of the data nodes fail (or be brought off line for backup). File Placement Optimization (FPO) is a technique to improve IO performance and with HDFS extensions allows for a compatible alternative to the HDFS in a Hadoop ecosystem.

Website application serving should also be configured for high availability, using Web servers to spread the load and adapt traffic distribution patterns should one of the application servers be brought off line. Distributed data caching is another consideration to provide direct memory access to frequently referenced content (e.g. the metadata taxonomy, or recently received / high interest content).

It is challenging to manage content to preserve copyright and limited access on the Web. Content can be altered to add unique signatures associated with the person consuming it to track provenance should the content pop up on unwarranted sites. Similarly, unique signatures can be incorporated during the initial registration process as artifacts become cataloged and represented electronically.

A balance must be struck between open access and protecting content. The repository should be aware of who is reviewing and who is receiving content, including their frequency visiting the site. Depending on agreements with education facilities or analytic service providers, it may be desirable to provide APIs that can assist with automating the access to content. Web and REST services are

a popular means to implement these APIs.

Data upkeep

As mentioned before, the content repository should be continually evolving with new information, insight, and deliberation about its artifacts. Newly discovered artifacts should be made available as quickly as practical with additional classification and information added as they become available. Analytics used to derive metadata should be run against all content to add perspective. These may need to be scheduled to run at off-peak usage periods, or may be run on separate systems to preserve performance goals for user interactions.

The user community should be consulted for feedback about the site, as they will provide many valuable suggestions to improve data organization and access. There should be people allocated to respond to the feedback and suggestions to ensure people's comments are addressed.

Possibilities

The potential benefit of having a federated digital humanities repository goes well beyond simply having one place to go for whatever data or artifacts one needs to find. Application of semantic metadata search, text analytics, collaborative deliberation and decision making, and cognitive computing capabilities can make it possible to integrate this information such that the system itself can draw inferences and make connections across disparate data. This makes it possible not just for researchers to access specific data and contribute their own analyses and interpretations, but also for the system to act as a research assistant, finding related data, artifacts, and commentary that the researcher didn't even know existed, and organizing and presenting it visually in a comprehensible manner. Combining the notions of repository, automated cognitive agent, and collaboration platform will enable the system to track users' expertise and interests, alerting them to relevant developments, and promoting a degree of collaboration and interaction between scholars that was not possible before.

At the time of writing, the Dutch humanities research community is awaiting the outcome of a bid to the Netherlands Organization for Scientific Research (www.nwo.nl) to develop a Common Lab for Research in the Arts and Humanities (CLARIAH). This 'Common Lab' will provide a sustainable research environment, which will provide researchers and research groups with integrated access to unprecedented collections of seamlessly interoperating computational research resources and innovative instruments to process them in virtual workspaces. The Common Lab is virtual, i.e. the data, instruments and facilities,

as well as its developers and users are distributed over various locations and institutes. The Common Lab will provide researchers with a wide variety of resources and services, e.g., intelligent access methods for exploring resources and innovative ways of combining different resources into virtual collections, so that information hidden in unstructured textual and multimedia documents, in combination with structured databases with qualitative and quantitative information, can be disclosed and analyzed. Interoperability of resources and services is a key element in the infrastructure, and thus it goes further than Europeana (www.europeana.eu) which provides a single access point for millions of digital cultural heritage objects. The infrastructure will be easy to access and use for scholars with limited technical training. Through dissemination activities, educational programs and training sessions, a new generation of researchers and students will be able to acquaint themselves with new research methodologies, thus creating the potential for groundbreaking research.

CLARIAH builds upon two large research facilities in the domain of the humanities on the 2008 Dutch Roadmap: CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities). CLARIN received funding for a national project (CLARIN-NL) at the Roadmap update in 2008. CLARIAH differs from CLARIN-NL, which is focused on language, in two respects: (1) it joins forces with the Dutch DARIAH community and its infrastructural needs, especially in the area of data and instruments for structured historical data, and (2) it deals with aspects hardly covered in CLARIN-NL, in particular audiovisual data and instruments, and facilities for virtual workspaces.

SOCIAL INFRASTRUCTURE

In this sub-section, we explore challenges in the social infrastructure for humanities research and collaboration – among scholars and with potential new collaborators among the public. Where appropriate, we highlight areas in which social and natural sciences may make contributions to these emergent humanities problems.

In this section we use the term ‘digital humanities’ (see note 1) in order to capture collaborative teams of researchers from different backgrounds. Many digital humanities (DH) projects require contributions from at least one humanities specialization and at least one computing specialization; some DH projects require more than one of each type of specialization; others require additional specializations, such as the social sciences for understanding collaboration or public engagement. We focus on how these issues may directly affect the work of DH scholars.

DH Needs and Current Technologies

We see three types of distance that DH scholars may encounter and that must be bridged. DH work may involve one or more of these distances:

- **Geographical distance.** While much humanities work has traditionally been done by individuals working alone or by co-located scholars, contemporary collaborations increasingly involve scholars working in different locations. The literature has documented how even a few hundred meters increases *collaboration difficulty*. Scholars working on different floors of the same building may find it difficult to collaborate, because of lack of opportunity for serendipitous meetings. These problems intensify with time zone differences. Computer and social science researchers have explored numerous instruments and ways of working that can decrease the problems of working remotely, such as shared online spaces and real-time communication. Beginning with a paper called 'Beyond being there,' other scholars have proposed that the goal of new instruments to support remote collaboration should not be to try to approach face-to-face collaboration (Hollan, 1992), but rather to take advantage of remoteness for new ways of accomplishing work with gains in diversity of knowledge or efficiencies in well-coordinated round-the-clock productivity. These technologies may help to move the opportunities for serendipity from face-to-face encounters to online encounters. The digital basis of new humanities research will ease the adoption of some of these technologies and strategies.
- **Temporal distance.** The humanities have always been concerned with contemporary and historical sources. Structure may be as significant as content (corpora, databases, etc.) in the creation of large-scale resources. Unlike private research notes and drafts, these resources are often intended to be used by others, long after their creation. Designing a representation for use by self or others in a longer-term future requires careful thought about how to (re-)approach the work when context and timeliness have both been lost. It may become necessary to capture not only data and structure within databases, but also the contemporaneous context (dictionaries, timelines, other databases) that contributed meaning to the primary databases.

  Unlike the other two dimensions, the temporal dimension also extends backward in time. That is, scholars must contend with how their work will be used in the future (as in the preceding paragraph), but also how their views (and the views of other scholars) are affected by the passage of time since an historical event or the creation of a literary or artistic work. Whose context is important? How does context affect interpretation? (Konst, Leemans and Noak, 2009) As historians of science have shown, meanings

change over time, and in some cases, 'facts' can also change over the timeframes envisioned by humanities scholars. Library and Information Science has contended with these issues, and has become one of the contributory disciplines to using computers within the humanities. Although computer scientists have been concerned with the creation and maintenance of reusable resources (principally code, but also data), their sense of 'timeframe' is orders of magnitude shorter than some of the humanities. Collaborations between humanities scholars, computer scientists, and social scientists could be of mutual benefit.

- **Disciplinary distance.** A powerful influence on collaboration is the need to cross disciplinary boundaries. It is obvious that different disciplines have their own departmental structures, career paths, and politics. In addition, different disciplines recognize different types of events and contributions as data (rules of evidence), and apply different logics regarding the 'right' way to work with those data (rules of inference). Computer and social scientists (especially anthropologists) have wrestled with these kind of difference in boundary-spanning encounters, such as those between engineers, scientists, designers, users, and people affected by the users. Those lessons can be applied as humanities moves into ever-widening circles of collaboration with diverse disciplines.

So far, we have discussed collaborations among scholars of various disciplines. Some scholars have experimented with radical forms of distributed work, by bringing laypeople into the research process – typically as voluntary labor, to provide data, or to clean or to enter analog materials. The natural sciences have found enormous value in crowd-sourcing data collection, data coding, and data analysis, under the general framing of citizen scientists. Scholars in the digital humanities have shared early research stages with people who might be called citizen humanists or perhaps lay researchers, especially with regard to primary data collection or document transcription. One potential growth area – for both natural sciences and humanities – is to broaden the participation by lay people, by recognizing their capacities as interpreters and knowledge creators.

Working with massive datasets and databases requires new methods for assembling resources and aggregating separate and distinct databases. Adding collaborative data, comments, and mark-ups to these separate information stores adds both complexity and intellectual power to the representational challenges. The sciences have developed a range of solutions to these problems, ranging from distributed relational database architectures to more heterogeneous 'federated database' architectures, to the 'network of networks' concept of the Internet itself. Many solutions from computing are ready for adoption.

Lessons learned from the more difficult integration/ aggregation projects can jump-start the combination of heterogeneous humanities resources.

Challenges

In contrast to massively distributed collaboration in 'big science,' the concept of 'big humanities' is only beginning to be envisioned. Humanities scholars will need to develop their own practices for collaboration across distance, time, and discipline. Once those practices are understood, scholars will be in a better position to evaluate the collaborative tools and technologies from the sciences, and to develop new instruments to fill in the inevitable, discipline-dependent gaps among those tools.

The existence of multiple mark-up standards (see 'Text and Social Analytics' section) will require some means to integrate these different expressions, representations, and technologies. The use of 'universal' terms may require shared dictionaries to resolve them, similar to the encoding standards in biology and astrophysics. For example, terms that have contextualized meanings may need some form of normalized representation, such as person names, place names, and the definition of relative time concepts such as 'after dinner' or 'at the beginning of autumn.' Meanings for these will change in different regions, centuries, cultures, and discourse traditions. Will scholars need to resolve these ambiguities through linked references to a shared 'person server,' 'location server,' or more challengingly, 'culture server'?

An additional challenge may occur as scholars from the different disciplines share their methods. Visualizations are increasingly used for communicating humanities findings. Drucker (2011) makes a subtler, disciplinary argument about the need to preserve the interpretative nature of the humanities as humanities scholars move into more computational arenas. Her argument may lead new visual methods (and perhaps new metrics behind the visual methods) to express those more *located* or *situated* (i.e., perspective-based) experiences and subjectivities (Harding, 2005).

Possibilities

We envision a set of collaboration platforms to enable work across the three types of distance discussed above. These platforms will provide assistance with coding, mark-up, and contextualization of humanities questions and contributions, first among humanities scholars, then including their students, and finally engaging the broader lay public. While the natural sciences have made important strides in many of these areas, the humanities need to understand their own distinct needs, create their own practices, critically evaluate

solutions from other disciplines, adapt selected solutions to some problems, and create their own humanities-informed solutions for other problems. To do this, we recommend that the 'Discipline' dimension be broadened to include computer scientists and social scientists, and possibly the lay public, with the goal of strengthening knowledge in all of the component disciplines.

These collaboration platforms can go beyond providing computer-mediated support for traditional collaboration patterns, and make possible new forms of collaboration that represent entirely new ways of working together. For example, we can go beyond working and collaboration in private to allow rich public collaborations where data, hypotheses, arguments, and analyses (both visual and textual) are shared, models of scholars' interests and expertise are maintained and used to inform and engage them in work going on elsewhere, and contributions from all over the world and from many disciplines and perspectives can be automatically aggregated and organized.

EDUCATION AND TRAINING

This White Paper has emphasized the use of computational technologies to enhance research capabilities. There are also two significant learning and educational issues. First, as humanities research comes to depend more heavily on computer technology, practitioners will need to master new skills such as web programming and using sophisticated analytics instruments. Second, there is an opportunity to transform humanities education by incorporating computational technologies into arts and humanities education at all levels. The current climate of excitement and innovation around online learning has implications for both.

Digital humanities training

The World Wide Web has become central to social, organizational, governmental and business life. Thus, understanding the Web from both a social and technical perspective is essential to a wide variety of disciplines and careers. To prepare humanities students and researchers to deal with digital data sources and computational instruments, four main training programs are already being offered or are in preparation by the universities involved in CHAT:

1. *Web Science Minor at VU:* A minor program for bachelor students. Courses include Web Technology, Web Analysis, Web Society, Web Science
2. *Training program in Digital Humanities 'crash course' at UvA:* Short condensed

program, including five modules, each of one day, e.g. Gathering Data, Data Capta, Preparing Data, Telling Data, and Cooking Data.
3. *Undergraduate program Media & Information at UvA*: An undergraduate program, composed from a mix of existing courses, media and culture (modular) courses, courses from the Digital Humanities minor and new courses.
4. *Joint VU-UvA Digital Humanities minor*: A minor program for bachelor students, including courses 'Introduction to eHumanities', 'From Objects to Data', 'Coding the Humanities', Methods Lab'.

1. Web Science Minor at VU: This focuses on the foundations of the Web, and provides a broad understanding of its intricate workings. Students explore the effects of the Web on society, from the flow of information to the social implications of a connected world. Courses provide students with a set of powerful instruments to research the Web and its effects on the world from different fields of interest. Whether you are a law student interested in jurisdiction in online environments, a social scientist interested in the social effect of the web, or a geologist interested in the ramifications of the Web's infrastructure on the physical world, the instruments and techniques you need to study the Web are found here.

2. Training program in Digital Humanities 'crash course' UvA: (1) *Gathering Data* – After a general introduction into the goals of Digital Humanities, we discuss the shift in perspective when we look at our objects of study as data points. What questions and methodologies does this new point of view prompt? We focus on the process of acquiring data and on how to structure them in a meaningful way, with APIs and writing REST queries to acquire data. (2) *Data Capta* – An introduction to the use of computational instruments in historical disciplines is followed by a hands-on session to get acquainted with the open-access Geographical Information System QGIS, using historical maps of the *Atlas der Neederlanden*. After a brief explanation of basic terms used in network visualizations and a demonstration of some examples, we experiment with the visualization tool Gephi. (3) *Preparing Data* – We discuss how to parse, filter and extract information from datasets. There are several steps needed to transform data, before they can be queried, represented and visualized. We use Google Refine and Gephi to explore datasets. (4) *Telling Data* – This session focuses on analysis of digital text, focusing on regular expressions as a more general and even more powerful way to search for patterns in text data. (5) *Cooking Data* – Focuses on presenting and interacting with data through visualization to present optimally the story we want to tell with the data.

3. Undergraduate program Media & Information at UvA: This is composed of a mix of existing courses, media and culture (modular) courses, courses from the Digital Humanities minor, and new courses. The existing courses include 'History of Information', 'Philosophy of Science', 'Research Working Group', and 'Bachelors Thesis'. The Digital Humanities minor includes courses 'Introduction to eHumanities', 'Methods Lab', 'From Objects to Data', 'Coding the Humanities'. The Media and Culture courses include 'Introduction Media culture I & II', 'Media History', 'Media Theory and Scientific Writing', 'Media culture in transformation', and 'Medialab'. New courses include 'Archives, Repositories and Curation', 'Search', 'Ubiquitous information and its consequences'.

4. Joint UvA and VU eHumanities Minor: This offers students insights into the meaning and value of computational approaches in humanities research. Data in the humanities, both digitized and 'digital born' are multimedial and rich. Humanities students in the digital age learn (1) how to cope with these 'new' forms of information, (2) what kind of choices are made in digitization processes, (3) which analytical instruments are at hand to analyze large amounts of complex data. It is crucial that humanities students learn to explore the ways computational methods and techniques influence humanities research. The minor closes with a 'Course Project', in which teams of three or four students collaborate with researchers and organizations in the field of eHumanities in small research projects. In these 'collaboratories', the insights and skills gained in the minor are used in a practical context, and the project teams present their results in a general workshop or seminar.

4.1. Introduction to eHumanities: This course introduces students to computational approaches in humanities research. Objects in the humanities, such as literary texts, historical sources, paintings or films, are often multimedial and rich. The course gives an overview of the field of ehumanities and provides students with knowledge of digitized and born-digital material: how do we cope with and interpret the 'new' forms of information, and which instruments are at hand to analyze and identify patterns in large amounts of complex data? The focus is on text, image and sound in digital heritage. Students engage in a critical reflection on the instruments and methods and explore the way computational techniques influence current humanities research.

4.2. From objects to data: This course considers a number of relevant

aspects for making data from objects, e.g. (1) exemplifying how to move from objects as objects to objects as data points, and how to use data to construct narratives, and research workflows; (2) dealing with availability of data sources, relevance of (meta)data, obtaining data from sources, data dumps, APIs, queries, and evaluating data; (3) working with structured and unstructured data, humanities material and metadata, parsing data, filtering data, segmenting and identifying relevant units in text and images; (4) mining for interesting data, building blocks for narratives, choosing a visual representation, refining data preparation, visual representation; (5) interaction, perspectives, scale, zooming, areas of interest for telling stories with data; (6) algorithms, repeatability, interpretation providing recipes and repetition, data interpretation and hermeneutics.

4.3 Source and data annotation: This course focuses on the process of annotation of sources and data and the implications for the theoretical models and concepts within different disciplines, by (1) annotating interdisciplinary datasets and using different computational instruments; (2) developing a code book and comparing the different annotations; (3) feeding a machine-learning program with the annotations they made and reflecting on the performance of the automatic annotation. The focus is on high-level semantic annotations that are of interest to a broader range of humanities and computer science students. Annotation is an important step towards the formalization of humanities: (1) it forces humanities researchers to represent their interpretation of sources in a data structure, and (2) it requires the use of some type of interpretation model and it results in an analysis that can be compared across annotators with different backgrounds, e.g. linguistics, (cultural-)historians, social scientists, literature specialists, and non-experts who may consider sources and data differently and thus arrive at different annotations of the same source/data.

4.4 Coding the Humanities: This course teaches a higher level, dynamically typed programming language, suited to the needs of the humanities (e.g. JavaScript) and various research techniques, e.g. imaging, translation, mapping, visualization, filtering, searching; publishing, linking; crowdsourcing, annotation. At the moment, there are no broadly available academic programming courses aimed at humanities scholars. However, coding skills are needed more now than ever, to help students and researchers to understand the various technologically mediated objects they study and to develop custom instruments that can improve the practice of humanities

research and the quantity and quality of its output. In this way, new forms of collaboration could be supported, e.g. humanities scholars could combine their different skill sets for larger, interdisciplinary projects, and more easily communicate with computer scientists and technical experts. Moreover, they can develop algorithmic instruments specifically for humanities scholarship and meet the ever-growing demand in the public and private sector for qualified people who can think computationally.

Both university partners via their information and computer science departments offer specialized courses in text and data mining, search engines, semantic web, machine learning, and intelligent systems, and the UvA offers a minor in programming that is open to all students (http://studiegids.uva.nl/web/uva/sgs/nl/p/1228.html).

Trends in online learning

We are in the midst of a rush towards online learning, led by the popularity (particularly in the U.S., but growing worldwide) of Massive Open Online Courses (MOOCs) such as Coursera (https://www.coursera.org). At the university level they present a new way of sourcing and sharing material, with an openness and scale far beyond past open universities. But their full potential is most likely to be a 'classic' disruptive technology story (Christensen et al, 2010): the underserved learner populations will now be served by a new technology, and their needs (rather than those of the entrenched population) will drive new requirements. The needs of lifelong learners and other non-standard student populations will dominate as the technology and services offered evolve over time.

MOOCs today are not very different from lectures, but now it is the massively distributed student body that will drive requirements. They may not want to learn in a prescribed order, so will browse and jump around and discuss material with others in online forums, critique the material openly, and will readily 'drop out' and go elsewhere if their needs are not being met. These same technologies that could distract from a crafted syllabus will also open up new opportunities to engage learners in online activities designed to complement the lectures. The technology provides more opportunities for engaging collaborative exercises, whether in the form of games, interactive visualizations, or other analytics instruments. MOOCs open up new opportunities for social learning, a range of collaborative methods for teaching and learning. We define social learning as learning done collaboratively with the support of others found through social networks. Today, nearly all online activity can contribute to lifelong, situated and social learning. From

serendipitous activity such as following someone on Facebook, one can keep up with new information through links, reading and assessing relative quality, 'liking' and sharing information. For targeted learning – researching a specific topic – one can find relevant content and colleagues, and then piece together material and conversations to master something new.

Coursera is an education company that partners with the top universities and organizations in the world to offer courses online for anyone to take, for free. Courses are taught by world-class professors and students join a global community of thousands of students learning online. A wide range of courses are offered, among others, in humanities, social sciences, and computer science, e.g. Computational Methods for Data Analysis, Information theory.

VideoLectures.NET (http://videolectures.net) is a free and open access educational video lectures repository. The lectures are given by distinguished scholars and scientists at the most important and prominent events like conferences, summer schools, workshops and science promotional events from many fields. The portal is aimed at promoting science, exchanging ideas and fostering knowledge sharing by providing high quality content not only to the scientific community but also to the general public. All lectures, accompanying documents, information and links are systematically selected and classified through the editorial process taking into account also users' comments.

With MOOCs, educational materials can be made more widely available, not just for those going to school, but for anyone interested in lifelong learning. The emerging social learning platforms provide a strong basis on which CHAT can build in order to integrate research with local cultural heritage institutions,

Learning challenges

As stated earlier, humanities research based on computational technologies will require algorithmic thinking skills. For some established researchers, this will require additional training that will need to be woven into the conventional career path. For both established humanities researchers and people in secondary school, university and doctoral programs, an emphasis on technology during training could be off-putting to those who chose humanities as a non-technical field. The value of the algorithmic instruments will have to be

emphasized, and presented in convincing ways.

Although MOOCs reach massive numbers of students, no viable business model exists, nor does an adequate, robust system of assessment. So while MOOCs could offer a platform for outreach, there are challenges around funding and sustaining MOOC-based training initiatives.

Online learning opportunities

The first opportunity is to use MOOCs for retraining humanities professionals. As noted above, there are challenges that would have to be addressed. But there are many offerings that touch already on the skills needed – big data analysis, visualization, statistics, programming – and can be leveraged without creating new courses. CHAT could participate explicitly by providing supporting online discussions, access to their data and computational instruments, and guest lectures at ongoing MOOCs.

Humanities content could itself be the basis for unique new offerings by CHAT. MOOC offerings are still primarily technical, so that there is an opening to provide new ones. Developing a course that integrated humanities content with exercises based on CHAT technology would be innovative, useful, and strong. Integrating mastery of some of the basic computing skills would also mean that these skills were introduced in context, rather than by sending humanities students to the computer science department, as many of their offerings might be too abstract and too far removed from the goals of humanities students. Similarly, CHAT could participate in a growing movement among computer science educators to make their field more relevant by providing exercises that could be used in computer science courses to make those courses more relevant to wider populations.

Research advances in humanities can become powerful learning tools. The availability of the research instruments for optional incorporation into curricula – or just for fun on their own – could be used to keep humanities prominently on the agenda of educators and policy makers. With so much concern for building STEM (science, technology, engineering and medicine) skills in schools, there is a danger that humanities education will be shortchanged.

An interesting business case could be made around lowered costs of MOOCs over on-campus learning. While this may not be a global trend, threats to liberal arts education in the U.S. exist. For example, the state of Florida is considering denying in-state tuition waivers to humanities majors on the grounds that they will not contribute as much to society as people trained in more practical fields. While we deplore this trend, and the related cost-cutting measures to primary and secondary education that eliminate arts programs, developing high-quality, lower-cost alternatives could be a way to assure that the benefits of humanities

education are still available to all.

Social learning tools that will engage and entertain will be particularly useful for promoting humanities. Collaborative projects such as crowdsourcing the reconstruction of historical events, personal and group genealogy, word analysis as in ManyEyes should be a natural product of a humanities research program.

CONCLUSION

In this White Paper, we have described important challenges in humanities research and important enabling technologies that would transform such research. As such, we believe that this White Paper may serve as the foundation of an interdisciplinary program of research to be initiated and supported by a new Center for Humanities and Technology (CHAT).

We believe that the research program of CHAT will provide opportunities for breakthrough humanities research, built on world-class technology and collaboration among an internationally recognized and diverse group of researchers. The inherently multidisciplinary approach will bring multiple perspectives to the research program and allow innovative programs to be executed.

The research program of CHAT will also be important for the fields of information and computer science. Historical and extremely heterogeneous data will require new analytic approaches and instruments. The content domains will push the boundaries of cognitive computing.

The CHAT research program will no doubt result in new research practices and paradigms for the humanities. These practices will serve as exemplars for digital humanities curricula as they continue to grow and evolve.

Finally, there is much promise in the societal benefits of the CHAT research program. New ways to explore and understand history, literature, and the arts will increase cultural understanding and promote greater social cohesion in an increasingly diverse and interconnected world. New algorithmic instruments and tools to enable understanding from *multiple perspectives* will enable a deeper and more reflective experience of human events.

REFERENCES

Aroyo, L, &. Welty, C. (2013). Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Web Science'13*, ACM.

Bod, R. (2013). *A New History of the Humanities. The Search for Principles and Patterns from Antiquity to the Present*. Oxford: Oxford University Press.

Borgmann, C. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly 3*(4).

Bowker, G., & Star, S.L. (1999). *Sorting things out. Classification and its consequences*. Cambridge MA: The MIT Press.

Bron, M., Balog, K., & de Rijke, M. (2013). Example based entity search in the web of data. In *Advances in Information Retrieval*, pp. 392–403, Springer.

Bron, M., Huurnink, B., & de Rijke, M. (2011). Linking archives using document enrichment and term selection. In *Research and Advanced Technology for Digital Libraries*, pp. 360–371, Springer.

Bron, M., Van Gorp, J., Nack, F., de Rijke, M., Vishneuski, A. & de Leeuw, S. (2012). A subjunctive exploratory search interface to support media studies researchers. In *SIGIR'12*, pp. 425–434, ACM, 2012.

Bron, M., Van Gorp, J., Nack, F., Baltussen, L.B. & de Rijke, M. (2013). Aggregated search interface preferences in multi-session search tasks. In *SIGIR'13*, pp.123–132, ACM.

Bulger, M., Meyer, E.T., de la Flor, G., Terras, M., Wyatt, S., Jirotka, M., Eccles, K., & Madsen, C. (2011). *Reinventing research? Information practices in the humanities*. London: Research Information Network.

Chang, D., Ge, Y., Song, S., Coleman, N., Christensen, J., & Heer, J. (2009). *Visualizing the Republic of Letters*. Stanford: Stanford University.

Christensen, C., Horn, M., and Johnson, C. (2010). *Disrupting Class. How Disruptive Innovation will change the way the World Learns*. New York: McGraw-Hill.

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly 5*(1).

Friendly, M. (2006). A Brief History of Data Visualization. In *Handbook of Computational Statistics: Data Visualization*, Vol III, Heidelberg: Springer-Verlag.

Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Aroyo, L., & Schreiber, G. (2013). An evaluation of labelling-game data for video retrieval. In *Advances in Information Retrieval*, pp.50-61.

Gligorov, R., Hildebrand, M., van Ossenbruggen, J., Schreiber, G., & Aroyo, L. (2011). On the role of user-generated metadata in audio visual collections. In *K-Cap'11*, pp. 145-152.

Harding, S. (2005). Rethinking standpoint epistemology: What is 'strong objectivity?' In A.E. Cudd & R.O. Andreasen (eds.), *Feminist Theory: A Philosophical Anthology.* Oxford: Blackwell Publishing.

Hollan, J., & Stornetta, S. (1992). Beyond being there. *Proc. CHI 1992*, 119-125.

Karsdorp, F., & Bosch, A. van den (2013) Identifying motifs in folktales using topic models. In *Proceedings of the 22nd Annual Belgian-Dutch Conference on Machine Learning*, pp.41-49.

Konst, J., Leemans, I., & Noak, B. (eds) (2009). *Niederländisch-deutsche Kulturbeziehungen 1600-1830*. Göttingen: V&R unipress.

Maccatrozzo, V. (2012). Burst the Filter Bubble: Using Semantic Web to Enable Serendipity. In *ISWC'12 Doctoral Consortium*, Springer.

Mosteller , F., & Wallace, D.L. (1964/2008). *Inference and Disputed Authorship*. Chicago: University of Chicago Press.

Newman, D., & Block, S. (2006). Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper Tools *Journal of the American Society for Information Science and Technology*, 57(6): 753–767.

Ockeloen, N., Fokkens, A., ter Braake, S. & Vossen, P. (2013). BiographyNet: Managing Provenance at multiple levels and from different perspectives. In: Blomqvist, E. & Groza, T. (eds), *Proceedings of the Linked Science workshop at the 12th International Semantic Web Conference and the 1st Australasian Semantic Web Conference* (ISWC2013), 21-25 October 2013, Sydney, Australia.

Ray, B. (2002) http://salem.lib.virginia.edu/maps/accusationMaps/townmap/trimap68.swf. The University of Virginia.

Roorda, D., Bos, E-J., & van den Heuvel, C., (2010). Letters, Ideas and Information Technology. Using digital corpora of letters to disclose the circulation of knowledge in the 17th century. In *Digital Humanities Conference Abstracts King's College London* (pp. 211-214).

Venuti, L. (2013). *Translation changes Everything. Theory and Practice*. London: Routledge.

Wang, Y., Wang, S., Stash, N., Aroyo, L., & Schreiber, G. (2010). Enhancing content-based recommendation with the task model of classification. In *Knowledge Engineering and Management by the Masses*, pp.431-440.

Wouters, P., Beaulieu, A., Scharnhorst, A., & Wyatt, S. (eds) (2013). *Virtual Knowledge*, *Experimenting in the Humanities and the Social Sciences*, Cambridge, MA: The MIT Press.

APPENDIX: COMPUTATIONAL METHODS IN THE HUMANITIES

Develop computational methods for the analysis of different perspectives, interpretations, and narratives on/about events and concepts, considering provenance and temporal aspects in collections of text, structured data and audiovisual material. There are three levels of operation: (1) digital analysis and representation of concept formation and the various (often conflicting) perspectives on these concepts; (2) illustration of how different analytical tools and methods lead to different interpretations; (3) evaluation of the knowledge process – valuation of different interpretations

Examples of Use Cases	
Art History *Understanding the perceptions of artwork, multi-perspective views of visual media, associations in different contexts*	• Can we detect meaningful relationships between artworks when we do not understand the semantic labels (due to language differences), or with insufficient clues (untitled works)? • Can we search for artworks on the basis of pattern recognition of e.g. color, composition, texture, rhythm?
Emotional Communities *Understanding emotional concepts, behavior, and embodied emotions in visual depictions, vocabularies and narrative structures*	• How to model and represent emotion detection, transmission interpretation (e.g. over time)? • How to understand the expression, conceptualization and transmission of feelings by groups/nations or location? • What is the drive for the emotional economy for user perception? • How do 'emotional communities' form specific perspectives on concepts?
Historical sites, bodies & objects *Understanding the shift in perspectives on heritage (narratives), e.g. from national to transnational, from European to global*	• How can we understand the epistemic communities that maintain heritage sites? • Can we represent and analyze conflicting historical narratives? • How can we help users with dynamically changing point of views to understand and make use of different narratives around heritage? • Can we trace how Europe disappeared from the national historiography of empires and how it has no contours in global heritage perspectives?
History of literature, language, law & politics *Understanding differences in perception, in culture, in temporal periods*	• Can we (semi-)automatically generate book narratives? • What do different linguistic expressions in books tell us about different perceptions hereof? • What are the similarities and differences of narrative structures in historical accounts, literary texts, video games representing a book?
History of Medicine *Understanding the impact of medical knowledge transfer*	• Can we detect changes in the perception of medical care in the exchanges of medical knowledge between the Netherlands and the Far East? • What are the differences in the perceptions of medical care for pharmacists, midwives by different social groups of patients? • What can the 'Prize Letters' tell us about the exchange of knowledge on medical care between the West and the East in the 17-19th Century?
History of Science *Understanding the cultural differences in the perceptions of the universe*	• How is the universe perceived in various cultures? • Can we trace signs of heliocentrism before Copernicus? • Is there an influence of Copernicus's writings on heliocentric views in other cultures?
Creative Cities *Understanding the cultural success of cities in terms of factors, dependencies, causality and perspectives*	• What are the factors that make a city a creative center of innovation? • Where are creative entrepreneurs located? • How do they communicate, interact, collaborate, and compete? • How do they turn the city into a magnet for other innovators?
Data & Sources	
Textual mentions, Named entities, Temporal and Location observations, User tags, Object metadata	

David Millen is a Senior Research Scientist at the T J Watson Research Center (Cambridge) and the Center for Social Business. His research interests include Social Computing, Human-Computer Interaction, and Computer-Supported Cooperative Work. In 2011, David was recognized by ACM as a Distinguished Scientist for his work in the area of social computing.

Sally Wyatt is Program Leader of the eHumanities Group of the KNAW, and Professor of Digital Cultures in Development at Maastricht University. Her research focuses on what digital technologies mean for humanities and social sciences research practices, and on the use of digital technologies in healthcare.

Lora Aroyo is Associate Professor at Web & Media group, VU University Amsterdam. Her research focuses on semantic search, recommendation systems, and event-driven access to online multimedia collections in the domains of cultural heritage and interactive TV. In collaboration with the IBM Watson group, she also works on crowdsourcing data collection methods for the adaptation of Watson system to the medical domain.

Rens Bod is Professor of Computational and Digital Humanities at the University of Amsterdam and Director of the Center for Digital Humanities. He has published over 200 articles and books, including *A New History of the Humanities* (Oxford 2013*), The Forgotten Sciences* (Amsterdam 2010, translated into four languages), *Probabilistic Linguistics* (Cambridge, MA 2003) and *Data-Oriented Parsing* (Chicago 2003).

Antal van den Bosch was trained as a computational linguist, and has worked in experimental psychology and computer science labs. He is now Professor of Language and Speech Technology at Radboud University Nijmegen. His areas of interest are machine learning of natural language, text analytics of historical and present-day big data text collections, and spelling and writing tools.

Irene Greif is a research scientist known for her contributions in the field of social and computer science. She founded the field of CSCW (Computer-Supported Cooperative Work), created and led the Collaborative User Experience Research Group at Lotus Development and then IBM. She has held the positions of Fellow and Chief Scientist at IBM, is a Fellow of ACM and AAAS, and was awarded The Anita Borg Technical Leadership Award in 2012.

Antske Fokkens is a researcher in computational linguistics at the VU University Amsterdam and is affiliated with the Netherlands eScience Center as an external engineer. Her research addresses methodological questions in various domains of computational linguistics including grammar engineering and text mining for digital humanities.

Charles van den Heuvel leads the research group History of Science and Scholarship at the Huygens Institute for the History of the Netherlands (KNAW). He also holds the Chair in Digital Methods and Historical Disciplines at the University of Amsterdam.

Inger Leemans is Professor of Cultural History at VU University, where she coordinates the Master's program 'Culture & Power'. Her interest in cultural economics has resulted in research about censorship, journalism, literary criticism, and the literary 'bubble' that accompanied the financial bubble of 1720. She is one of the founders and directors of ACCESS - The Amsterdam Centre for Cross-Disciplinary Emotion and Sensory Studies.

Susan Legêne is Professor of Political History at VU University Amsterdam, specializing in colonialism, decolonization and post-colonial state formation. Trained as a historian, she also has a professional background in development cooperation and the museum sector. At the VU Research Centre CLUE, she is one of the leaders of the research cluster Global History and Heritage in a Post-Colonial World.

Mauro Martino is Research Manager at IBM's Center for Innovation in Visual Analytics (CIVA). His projects have been shown at international festivals and galleries, including Arts Electronica, Serpentine Gallery (London), and GAFTA (San Francisco). His work has been featured on the cover of *Nature*, and in *Popular Science, The Economist, The Financial Times, WIRED Magazine, The Guardian, BBC News, MIT News,* and *Harvard News*.

Nathaniel Mills has been a Senior Technical Staff Member at IBM Research since 1997. His specialties include complex systems management, collaborative decision making, and empirical, analytic, and semantic reasoning. He has won accomplishment awards for products relating to web performance, and analytics for condition based maintenance of complex systems.

Merry Morse is a member of the Collaborative User Experience Research group within IBM Watson Research. Her current focus is on social business and social learning and her background is in Human Computer Interaction, User Experience, and Education.

Michael Muller is an internationally recognized expert on participatory design and social media. He is an ACM Distinguished Scientist and an IBM Master Inventor.

Maarten de Rijke is Professor of Information Processing and Internet at the University of Amsterdam, and a *Pionier* grant laureate of the Netherlands Organization for Scientific Research. He heads one of the world's leading research groups in information retrieval and has a special interest in semantic search, social media analysis, and self-learning search engines.

Steven Ross is a Senior Technical Staff Member of the Collaborative User Experience group in IBM Research. His research is focused on Human Computer Interaction, Visualization, and Collective Intelligence.

Piek Vossen is Professor of Computational Lexicology at the VU University Amsterdam, and founder and president of the Global WordNet Association. He combines linguistics and computer sciences to analyze linguistic phenomena using computer models. His latest project is the 'History Recorder', a computer program that 'reads' the news each day and precisely records what happened when and where in the world and who was involved. In 2013 he was awarded the prestigious Spinoza Award.

Chris Welty is one of the twelve original members of the team that built Watson, the question-answering computer. He continues to lead the structured knowledge exploitation and rapid domain adaptation teams at IBM's Watson Technologies Division. He has published widely in the areas of Ontology, Semantic Web, and Natural Language Processing.