Accomplishments Section

*What are the major goals of the project? (8,000 characters)*

The major goals of the project were the following:

1. *Incentives and disincentives of scholarly collaborations:*
   a. Study how voluntarism organizes and structures academic discourse, with cyberinfrastructure allowing for an intellectual architecture that exists over, outside, beyond, and sometimes despite the silos and institutional structures of higher education.
   b. Assess the motivations for peer-networking and knowledge-sharing in online spaces, especially for early-career professionals for whom the rewards of such extra-institutional collaboration might seem unclear,
   c. Ask how virtual peer-mentorship contributes to career development, including risk taking,
   d. Analyze data on 868 current students and recent alumni (80% graduate and 20% undergraduate students) in the HASTAC Scholars program in order to determine whether individuals in HASTAC’s networked interdisciplinary organization do or do not "outcompete" those young professionals whose work remains fixed in more traditional hierarchical, linear disciplinary professional structures.
   e. Address whether there is a difference in the success rate of underrepresented minorities with this collaborative, peer-mentoring system.
   f. Analyze what factors incentivize or disincentivize collaboration, how discourse on the site translates into cyberinfrastructural transformation (such as institutional
impacts, pedagogical applications, research impacts, or tool management or deployment).

2. **Data extraction and analysis:**
   a. Data-mine the 176MB MySQL database on which the HASTAC.org site is based, assessing data, metadata, and taxonomies related to anonymized user profiles, content, and comments, and revealing connections across tens of thousands of individuals, institutions, professional associations, conferences, publications, research projects, grants, tools, and programs.
   b. Evaluate ties between individuals within (and potentially beyond) the 13,000-member network,
   c. Explore the connection between tags, topics, and content on the site.
   d. Apply computational text analysis to the textual content.

3. **Peer participation and feedback:**
   a. Post preliminary findings as ongoing public progress reports on the HASTAC site in order to invite comment, feedback, and participation.
   b. Provide better understanding of cost-benefit outcomes for academic organizations seeking the best ways to support professional development.

* What was accomplished under these goals?

**Major Activities (8,000 characters):**

1. **Incentives and disincentives of scholarly collaborations:**

We investigated the incentives and disincentives of scholarly collaborations across the HASTAC network. Because HASTAC is an open network, we have published our progress on this grant in a series of blog posts produced between January 18, 2013 and December 31, 2014. This material is based on research conducted by Postdoctoral Fellow David Sparks (January 2013 to August 2013) and then by Postdoctoral Fellow Marco Toledo Bastos (October 2013 to December 2014). These public blog posts can be accessed at [http://www.hastac.org/collections/eager](http://www.hastac.org/collections/eager).

The results reported on the blog posts explore the dynamics of HASTAC network, the motivations for peer-networking and knowledge-sharing in online spaces, the interplay between peer-mentorship and geography, and the factors that incentivize collaboration on online scholarly social networks. The preliminary findings were presented at the "7th International Conference of the Humanities, Arts Science and Technology Alliance & Collaboratory (HASTAC’14)”, the "Big Data & Collaboration Conference,” the "General Online Research 2014 (GOR’14),” at the "Digital Humanities Berlin (DHB’14),” and the 48th Hawaii International Conference on System Sciences (HICSS’15).
For the purposes of this project, we developed and implement a Qualtrics survey that measured the correlation between online collaboration and academic output. The survey collected data from 156 HASTAC Scholars that completed the questionnaire from a total of 868 Scholars that received the invitation (see "ANNEX I - HASTAC Scholars Survey Report.pdf"). Data collected with the survey were used to determine whether HASTAC users "outcompete" young professionals in more traditional, hierarchical, and linear disciplinary professional structures (see section "Significant results" for the results of this study).

These are core questions of the EAGER grant that were explored in the paper "Bridging Structural Holes: Scholarly Collaboration in Online Social Networks," presented at the 7th IEEE/ASE International Conference on Social Computing and made available at the ACM Digital Library. A lengthier and more detailed version of this study, also based on the results of the survey, was presented at the 48th Hawaii International Conference on System Sciences (HICSS’15). Both papers are dedicated to determining whether individuals in HASTAC’s networked interdisciplinary organization do or do not "outcompete" young professionals in more traditional hierarchical, linear disciplinary professional structures (see section "Significant results" for the results of this study).

2. Data extraction and analysis:

The HASTAC network comprises a 176MB MySQL database that includes the HASTAC.org site. We mined this database to assess the data, metadata, and taxonomies related to anonymized user profiles, content, and comments. The results reveal of this exploratory analysis reveal connections across tens of thousands of individuals, institutions, professional associations, conferences, publications, research projects, grants, tools, and programs. We evaluated the ties between nodes within and beyond the network and explored the connection between tags, topics, and content on the site. The results were reported on a series of blog posts covering the HASTAC EAGER research.

3. Peer participation and feedback:

The reason for publishing our progress on this grant in a series of blog posts on HASTAC website was to invite comment, feedback, and participation. The research team also created a GitHub repository for material supported by the EAGER grant that is open for public access and citable with a Digital Object Identifier. In order to assess the reach of HASTAC content across social networks, we collected the number of hits each post published on HASTAC.org received across a number of social networks. The report on HASTAC outreach includes the number of Facebook comments, likes, shares, and clicks; the number of tweets on Twitter, pins on Pinterest, views on StumbleUpon, shares on LinkedIn; and the number of ups, downs, comments, and the score on Reddit for every post published on HASTAC.org since 2006 (see "ANNEX II - HASTAC Social Media Report.pdf").
Specific Objectives (8,000 characters):

The specific objectives of this project were deemed accomplished and fall into three categories: data mining, hypothesis testing, and information dissemination.

**Data Mining:** our first objective was to extract and mine the SQL database containing the structure and content of the HASTAC.org website. This was accomplished within the first week of our postdoctoral fellow's tenure on staff and soon thereafter the process was streamlined by the new postdoctoral fellow's tenure. During this second phase of the project the SQL database was loaded directly onto R, which facilitated the process of cleaning and organizing the database. Essentially, most of the tables in the database referred to nodes, users, tags, and topics assigned to each of the posts published on the website, while the bulk of the data, in terms of file size, consisted of textual content. This textual content, in particular, proved time-consuming to refine, but such refinement was necessary for much of the work to follow.

**Hypothesis Testing:** we relied on our primary source of explanatory variables (the HASTAC.org database) to measure concepts like interconnectedness, degree of participation, and disciplinary breadth of research interests for HASTAC's members. Specifically, we were interested in understanding how online scholarly networks and involvement with the community can impact the careers of early-career scholars working with interdisciplinary research. To this end, we developed and deployed a survey instrument designed to tap HASTAC members' self-reported academic output and collaborations with other HASTAC members and beyond, as well as their engagement to the HASTAC network and the level of interdisciplinarity in their academic background.

We used the results of this survey to evaluate the extent to which HASTAC provided opportunities for collaboration, how invested members were in those partnerships, and the degree to which members' research interests were influenced by their involvement. We collected traditional and non-traditional measures of scholarly achievement, output, and cross-disciplinary integration and built a database of publication records, co-authorship, journal placement, and employment histories for all HASTAC members. This data allowed us to calculate authorship and co-authorship as a function of intra and interdisciplinary backgrounds.

The survey collected data on 123 students and recent alumni of the HASTAC Scholars Program (14% response rate). After processing the questionnaire responses, the data was collated together with HASTAC website data for hypothesis testing. We explored the network activity of users (including HASTAC Scholars) that reveal connections across thousands of individuals, institutions, professional associations, conferences, publications, blog posts, blog comments, forums, research projects, and grants. As a result, the survey study implemented with the Qualtrics platform and the data extract from the website allowed for testing the interplay between
users’ activity in the HASTAC network and academic output. To this end, we tested the following hypotheses:

H1. The activity level of HASTAC Scholars in the scholarly social network is associated with their academic output.
H2. HASTAC Scholars with interdisciplinary backgrounds present higher academic output.
H3. HASTAC Scholars with interdisciplinary backgrounds present higher activity level in the scholarly social network.

In a follow-up investigation we attempted to look at the relationship between online social networking activity and academic scholarship in a broader context by expanding the data analysis beyond the HASTAC network and including other online social networking sites to which researchers may have been engaging. To this end, we relied on the Hypotheses platform, another such online scholarly social network, that provides a means of comparison between multiple instances of networked scholarship. This study was published by the journal PLOS ONE and relied on HASTAC.org and Hypotheses.org data from two database dumps containing the SQL table structure and the blog post content. We filtered HASTAC’s and Hypotheses’s database for blog posts published in English between the 1st of July 2006 and the 30th of June 2012, thus extracting 7,269 posts from HASTAC and 6,783 posts from Hypotheses. After applying co-word analysis and topic modeling to the data and showing thematic similarities and differences between the two sites, we tested the following hypotheses:

H4. Digital Humanities content is dependent on their related traditional humanities discipline.
H5. The thematic differences in the distribution of topics in the two sites reflect different adoption of Digital Humanities terminologies in different national contexts.

**Information Dissemination:** Finally, we accomplished several specific objectives in the dissemination of our findings. Our primary mode of communication has been through the HASTAC EAGER Collection, a repository of blog posts associated with this project. Each of these blog posts highlighted a different stage of our research. When the exploratory period was completed and the material posted on the EAGER repository, the blog post was given prominent placement on the HASTAC.org homepage and publicized through the @HASTAC Twitter account. The HASTAC website sees approximately 35,000 visitors per month, many of whom are members and repeat visitors, so this repository of blog posts was a prime venue for reaching the most relevant audience.

We also worked very hard to develop visualizations of our findings, as an aid to communication with understanding. Visualization is a valuable tool for summarizing "Big Data" and complex phenomena, and our two postdoctoral fellows were well-versed in developing new modes of visualizations for distinct data challenges. These visualizations played a prominent role in the research presentations focusing on the work supported by the EAGER grant.
Finally, in addition to the seminars, workshops, and presentations associated with the grant we held a workshop on May 28, 2014 entitled "Big (and messy) Data & Collaboration Workshop & Conference." This one-day workshop was sponsored by HASTAC, the NSF EAGER Grant team, and the Duke University PhD Lab on Digital Knowledge. We invited researchers interested in the impact of scholarly networks to cross-disciplinary, multi-institutional research to discuss the analysis of big (and sometimes messy) data in academic, collaborative settings. The event was very successful and brought together many collaboratories focused on interdisciplinary and cross-institutional research. Another outcome of the workshop was the creation of an online forum on Complex Data & Collaboration.

**Significant results (8,000 characters):**

The significant findings of this research are detailed in the papers 1. Outcompeting Traditional Peers? Scholarly Social Networks and Academic Output (HICSS’15); 2. Bridging Structural Holes: Scholarly Collaboration in Online Social Networks (SocialCom’14); 3. Scholarly Collaboration in the HASTAC Social Network (BigDataScience’14); 4. How Digital are the Digital Humanities? An Analysis of Two Scholarly Platforms (PLOS One); and 5. Bales of HASTAC: On the Character and Structure of an Online Collaborative Alliance. Below we briefly summarize the findings (unpublished).

**Scholarly networks associated with academic output:** We analyzed the relationship between online social networking activity and offline academic scholarship and found considerable cross-pollination between the two activities. Although the regression model used in our study explained at best only 40% of the variation in the HASTAC website, the results show that most scholarship published by HASTAC Scholars in the period took place in partnership with other scholars that are also very active in the online network (77%). The remaining collaborations (23%) took place mostly between users registered to the website (17%) and a minority (4%) of the co-authored scholarship published in the period happened with Scholars not registered to the network.

**Interdisciplinarity not associated with academic output or online activity:** Although the HASTAC Scholars Survey data are too limited to fully test the hypothesis that the intra or interdisciplinary background of Scholars is associated with academic output and/or the activity in scholarly networks, our results rejected the general hypothesis that Scholars engaged with interdisciplinary work present a higher output measured by academic or website activity (although Scholars dedicated to interdisciplinary work presented a higher level of co-authorship on academic works compared to peers with intradisciplinary background). Further research focused on the relationship between intra and interdisciplinary background and scholarly social networks.
networks is required to advance our understanding of social network growth, community formation, and learning development in scholarly contexts.

**HASTAC Scholars are committed to academic research:** Our analyses show that the potentially confounding variable academic activity is the driving force in our model. In the last instance, we submit that both academic output and activity levels in the HASTAC network are driven by Scholars committed to academic research and that rely on scholarly social networks to further strengthen their academic curriculum. In short, Scholars do not achieve higher academic output as a result of joining scholarly social networks. Rather, Scholars join academic social networks because they are committed to research and are already involved with learning activities supported by digital networks.

**Networked Scholars "outcompeting" more traditional peers:** our results have revealed that the level of engagement in scholarly networks is associated with academic output. This relationship suggests that networked young scholars can potentially "outcompete" more traditional peers that refrain from engaging in online scholarly networks, as most coauthored scholarship published by HASTAC Scholars in the period resulted from collaborations with other users of the network (77%). However, we would ultimately caution against overemphasizing the impact of social networks on academic output, and would rather emphasize the role played by macro-level interactions between university groups and micro-level variables associated with Scholars’ personal commitment to academic research.

**Digital Humanities depends on Humanities:** our analyses of the HASTAC and Hypotheses networks showed a substantial one-way dependency of Digital Humanities terms on their respective main humanities area. We found a much higher ratio of posts mentioning concepts in the traditional humanities compared to their digital counterparts, which is indicative of a one-way dependency of digital concepts on the corresponding traditional humanities area. The results also suggest that digital archives, digital literacy, and digital pedagogy are relatively independent from their corresponding traditional humanities discipline, and that digital publishing, digital libraries, and digital media show considerable cross-pollination between the specialization and the main humanities area.

**HASTAC is an interdisciplinary DH network:** we found that interdisciplinarity is a major strength of the HASTAC network, particularly of the HASTAC Scholars group. DH-related labels are considerably more frequent in HASTAC.org than in Hypotheses.org and these differences are likely to mirror the unequal adoption of Digital Humanities terminologies in different national contexts. We found a tendency in Hypotheses.org towards focused thematic areas representing disciplinary interests contrasted with a tendency to discuss more cross-disciplinary themes in HASTAC.org.
**Skewed Distributions:** We have found that, like many other distributed online phenomena, many aspects of HASTAC.org follow a power-law distribution. The number of contributions by each user, the number of comments on each blog post, the number of times a given tag is used, and the number of page views on a given date all follow this pattern. A small number of each (user, blog post, tag, date) are responsible for a large proportion of observations (contributions, comments, tag uses, views), but there are very many which each contribute a few. The implication of this reoccurring "long-tailed" distribution is that many aspects of HASTAC.org are poorly represented by measures of central tendency. We found that many of these aspects are also found in other scholarly social networks that emerged in different national contexts.

**Global visitors, local contributors:** Using a geolocation database, we were able to map a subset of visitors and contributors to HASTAC.org. Within North America, most traffic and content comes from individuals in urban centers, particularly towns and cities with major universities. Globally, we found that HASTAC.org content contributors are geographically concentrated in North America. The vast majority of comments to blog posts came from sources within the United States, and nearly all comments came from the "Global North" with only 204 comments from longitudes south of the equator. This distribution emphasizes the fact that HASTAC, while ostensibly a global community, is particularly absent from the "Global South." We did find, however, that visitors to the site—those who come to read and learn about ideas, research, conferences, and opportunities in digital humanities and education, even if they don’t author content on the site—derive from a much broader global range. To close this geographic divide, HASTAC has engaged with many innovators, academics, and researchers in the "Global South" via ties to the Organization of American States and decided to hold the 2014 HASTAC annual conference in Lima, Peru.

**Densely connected, clustered topics:** The network of user-assigned tags as applied to blog posts on HASTAC.org is very dense. There are a small number of tags that are used very frequently, and so no two terms are particularly distant in terms of spanning a network of co-occurrence. As such, we used a novel matrix algebraic transformation of this matrix to identify the degree to which each pair of tags was related by users, regardless of the frequency of those terms' usage. This log-odds ratio co-occurrence matrix enabled us to estimate a latent semantic space for all tags in use on HASTAC.org and identify clusters within this space. A subjective classification of spatial tag clusters suggests dense pockets of topical conversation around badges, the internet, and opportunities (calls for papers, conferences, open positions), with somewhat less tag density around learning and a very heterogeneous group termed cultural studies. Smaller clusters could also be found around digital humanities, digital media and learning, visualization, and data.

**Key outcomes or Other achievements (8,000 characters):**
Key outcomes and achievements fall into three categories: publications in conference proceedings, presentations at academic conferences, blog posts, and academic papers.

Conference Proceedings


Conferences Presentations

2. M.T. Bastos presented the paper "Bridging Structural Holes: Scholarly Collaboration in Online Social Networks" at the 7th IEEE/ASE International Conference on Social Computing in Beijing, China.
4. M.T. Bastos presented with Cornelius Puschmann the paper "Writing styles and knowledge dissemination in academic networks" at the 2014 General Online Research in Cologne, Germany.
6. M.T. Bastos presented "The HASTAC Network, Visualized" based on the work supported by the EAGER grant in the panel "What we Learn from HASTAC Growing Network" at the 7th International Conference of the Humanities, Arts Science and Technology Alliance & Collaboratory (HASTAC'14) in Lima, Peru.
7. M.T. Bastos presented "Modelling the HASTAC Network" based on the work supported by the EAGER grant at the Conference "Big Data & Collaboration" held at Duke University in Durham, NC, USA.
8. D. Sparks presented "Diving Deeply into HASTAC's Big Data" based on the work supported by the EAGER grant at the 6th International Conference of the Humanities, Arts Science and Technology Alliance & Collaboratory (HASTAC’13) in Toronto, Ontario, Canada.

Academic Papers


Software


Reports


Blog posts

The following blog posts were posted to the EAGER Collection at HASTAC.org providing a continuous overview of the work:

1. "Outcompeting Traditional Peers?," in which we comment on the papers presented at the SocialCom'14 and HICSS’15 conferences.
2. "Introducing SocialMediaMineR: a social media tool for R," in which we introduce the R package SocialMediaMineR used to create "The HASTAC Social Media Report."
3. "Why do data types matter?" in which we discuss ethical considerations of data extraction, methodological approaches, and research challenges stemming directly from the unprecedented scale of data generated by users.
6. "HASTAC Topics by U.S. State (2006-2013)," in which we mapped the geographic distribution of user-assigned tags and site-proscribed topics by U.S. State (from 2006 to 2013).
7. "Self-Loops and Network Awareness," in which we analyzed the network of hyperlinks on HASTAC.org.
8. "#FutureEd in Clouds: University, Humanities, and MOOCs," in which we visualized the classes and institutions affiliated with the FutureEd Initiative by modelling the keywords and plotting the results in a word cloud.

9. "Users Are from Mars, Commenters Are from Venus," in which we bridged my work with what David Sparks has done before in terms of geolocation and spatial analysis. For this post I created some cartograms that are very useful at explaining the geographic differences between HASTAC blog authors and commenters.

10. "HASTAC Scholars and the Geography of Topics," in which we spatially analyzed the topical interests of HASTAC Scholars.

11. "What are HASTAC Scholars 2013-2014 talking about?," in which we used the Scholars biography, project, and keywords to map the relationship between Scholars and topics.

12. "Where are 2013-2014 HASTAC Scholars coming from?," in which we created a rank of incoming HASTAC Scholars per state in America by normalizing the data by population density in the US.

13. "Less yack and more hack at #MozFest," in which we reported the Mozilla Festival I attended from October 25-27, in London, to represent HASTAC.

14. "From Russia With Computer Love," in which we reported on my trip to Russia to deliver public talks at the Higher School of Economics in St. Petersburg and Moscow.

15. "A Snapshot of the HASTAC Network," in which we graphed and visualized the network of blog posts and comments at HASTAC.org.

16. "The HASTAC Spectrum," in which we explored the members' self-identified "Interests," as listed in their profiles.

17. "DML & Scholars are at the heart of HASTAC," in which we explored the largest groups in the HASTAC network.


19. "Understanding 'Centrality Bias' in Teams," in which we commented on Jonathon Cummings' research on "Centrality Bias."

20. "The geography of HASTAC.org," in which we mapped the HASTAC commenter locations across the globe and make limited inferences about the whole population of HASTAC members.

21. "A pattern-finding HASTAC puzzle," in which we request input from the HASTAC community to help identifying topics grouped using topic modelling tools.

22. "EAGER to begin," in which we welcome the opportunity to study HASTAC.

Timeline

Note that the original timeline is based on the grant being awarded August 2012. The first EAGER post-doctoral fellow did not join the project until January 2013 and left on August 2014. The second EAGER post-doctoral fellow joined the project on October 2014.

- Year One: Discovery and Findings
○ Technology-Assisted Participation Working Group convened to help strategize the optimal ways for extracting and analyzing data.
○ Postdoctoral Fellow, David Sparks, hired and began working January 7, 2013.
○ Initial extraction of MySQL database completed January 2013; second-phase work is now underway identifying and characterizing connections across user-profiles, posts, topic tags, and institutions.
○ Textual and social network analysis of content, tags, and other metadata completed, and incorporated into a white paper, presented at the HASTAC 2013 conference.
○ Visualizations developed in parallel with statistical analysis, and incorporated into white paper.
○ Technology-Assisted Participation Working Group re-convened to provide feedback on work accomplished to date and to help strategize additional ways of extracting and analyzing data.
○ White paper on computational analysis, data extraction, and social networking analysis of the interplay of cyberinfrastructure and scholarly communication in an academic peer-produced network presented at Duke's John Hope Franklin Center's "Wednesday at the Center" and the HASTAC Conference in Toronto, ON (April 2013).
○ Designed survey on technology-aided participation and mentoring to measure relationships and interdisciplinarity among HASTAC Scholars, received IRB approval (June 2013).

● Year Two: Visualization, Publication, Policy
○ Postdoctoral Fellow, Marco Bastos, hired and began working October 1, 2013.
○ Full extraction of HASTAC's MySQL database directly into R followed by cleaning and organizing the database.
○ Mining and graphing HASTAC network relationships extracted from the SQL database, particularly: 1. network of blog posts and comments; 2. network of users and concepts; 3. affiliation network of HASTAC Scholars based on co-participation in forums; and 4. Network of users and geographic locations.
○ Restructuring and rewriting of the survey questionnaire. The new version of the survey is less focused on mentorship and was designed to 1. measure the correlation between online collaboration and academic output; 2. assess the perceptions of HASTAC Scholars about the process of peer-learning; 3. gather demographic information about the user base. The survey was intended to collect data on 800+ current students and recent alumni of the HASTAC Scholars program and the leading research question was to determine whether HASTAC users "outcompete" young professionals in more traditional, hierarchical, and linear disciplinary professional structures.
○ Importing new survey questionnaire to Qualtrics.
The changes made to the June 2013 version of the questionnaire were described and submitted as an amendment to the Survey protocol, IRB approval was granted on January 3, 2014.

Coordination and implementation of the HASTAC Scholars survey. The Qualtrics Survey was activated and the 866 HASTAC Scholars were notified on January 20. The survey run until April 21, when the results were collected.

The short-papers "Writing styles and knowledge dissemination in academic networks" and "What's in a digital humanities blog? A comparison of HASTAC and Hypotheses," coauthored by M.T. Bastos and C. Puschmann, were accepted and presented at the 2014 General Online Research in Cologne, Germany, and the 2014 Digital Humanities Berlin, in Berlin, Germany, in February and March 2014, respectively.

The paper "How Digital are the Digital Humanities? An Analysis of Two Scholarly Platforms," coauthored by M.T. Bastos and C. Puschmann, was completed, submitted, and published at the Journal PLOS ONE.

Organized the Conference “Big Data & Collaboration” on May 28, 2014 and invited scholars to present their research at the event. The Conference Website went online on April 24, 2014, and the conference was held at the Duke University PhD Lab on Digital Knowledge with the participation of 40+ attendees.

Data mining of HASTAC Scholars survey focusing on the effects of online scholarly networks to young scholar's academic output. The data allows for limited testing of the hypotheses that 1. networked interdisciplinary young scholars "outcompete" more traditional peers; and 2. Scholars engaged with interdisciplinary work present a higher activity level measured by academic or network output. The data was used for drafting the following papers authored by M.T. Bastos: "Bridging Structural Holes: Scholarly Collaboration in Online Social Networks," "Scholarly Collaboration in the HASTAC Social Network," and "Outcompeting Traditional Peers? Scholarly Social Networks and Academic Output."

Social network analysis of HASTAC users, content, and tags completed and incorporated into a white paper, presented at the HASTAC 2014 conference.

Modelling of HASTAC's social network using exponential-family random graph models (ERGM) was completed and incorporated into a white paper, presented at the Conference "Big Data & Collaboration" held at the Duke University PhD Lab on Digital Knowledge.

Completion of the "HASTAC Social Media Report (2006-2014)" that reports the number of hits that each of HASTAC article has achieved across various social networking sites.

As a result of the work for "The HASTAC Social Media Report (2006-2014)," an R package was created and submitted to CRAN on June 23. The package is called SocialMediaMineR and it retrieves information about the popularity of URLs on
Facebook, Twitter, Pinterest, StumbleUpon, LinkedIn, and Reddit. SocialMediaMineR was made available on CRAN on June 26, 2014.

○ The paper "Bridging Structural Holes: Scholarly Collaboration in Online Social Networks" authored by M.T. Bastos was submitted, presented, and later published on the Proceedings of the 7th IEEE/ASE International Conference on Social Computing, held in Beijing, China.

○ The poster "Scholarly Collaboration in the HASTAC Social Network" authored by M.T. Bastos was accepted and published on the Proceedings of the 2nd IEEE/ASE International Conference on Big Data Science and Computing, held in Beijing, China.

○ The Survey protocol was renewed for another six months. IRB approval was granted on June 11, 2014 with expiration date on June 13, 2015.

* What opportunities for training and professional development has the project provided? (8,000 characters)

This grant has supported the Postdoctoral Scholar David Sparks from January to August 2013. From October 2013 to December 2014 it supported the Postdoctoral Scholar Marco Bastos. Since his hire, Bastos has gained experience as a researcher/scholar, as a data scientist, as a valued member of a nonprofit, and as a communicator able to translate specialized research to a larger audience of educators. Bastos and previously Sparks have led the data extraction, analysis, and visualization for the project, producing blog posts, academic white papers, and conference presentations. This experience has strengthened their ability to use R to analyze and visualize complex data systems, and it has provided an opportunity to present their findings to a community different from the one they were previously accustomed to (an audience predominantly consisting of humanities and technology scholars, researchers, and professionals). Moreover, Bastos and Sparks have also gained experience in the functions of a nonprofit, attending weekly staff meetings and learning about organizational planning and leadership.

* How have the results been disseminated to communities of interest? (8,000 characters)

Substantial effort has gone into disseminating the results of this research to communities of interest. The primary community of interest, in this case, is the approximately 13,000-member alliance that forms HASTAC itself. HASTAC is the "exemplar academic network" named in the grant proposal, and serves as both the primary subject of, and audience for, the present inquiry. As specified in earlier reports, we have taken three approaches to dissemination: online, events, and conferences.

Online: first, and most crucially, each phase of the research project, as it has neared completion, has been shared directly with the HASTAC community through the EAGER Collection at HASTAC.org. The published blog posts cover, inter alia, the geography of HASTAC
contributions, the structure of the HASTAC network, and the nature of topics under discussion on the site (see details on section "blog posts" above). In several cases, the Postdoctoral Scholar Marco Bastos sought and received feedback which directly informed his later research -- HASTAC members themselves are subject-matter experts when it comes to their own interest and communication, and their input was critical to the project. In each case, these blog posts were publicized through the HASTAC Twitter account, which has more than 8,000 followers, with the intent of reaching an audience that would otherwise not be aware of these research activities.

**Events:** the other primary means of dissemination was through a plenary presentation at the 2013 and the 2014 HASTAC conferences. The 2013 conference was held at York University near Toronto and the 2014 conference was held at the Ministerio Cultura of Lima, in Peru. In each iteration the conference has brought more than 300 HASTAC members and contributors interested in sharing research with others across the interdisciplinary alliance. In 2013 in a panel entitled "HASTAC and Humanities in a Digital Age: The View Ten Years Out," Sparks presented slides summarizing the research project until that point in time. In 2014 in a panel entitled "What we Learn from HASTAC Growing Network" Bastos presented slides summarizing the EAGER research at the 7th International Conference of the Humanities, Arts Science and Technology Alliance & Collaboratory (HASTAC'14) in Lima, Peru. These presentations served to enhance interest in this project among individuals who are influential within HASTAC and in the broader academic and research communities. Additionally, Bastos also presented his findings based on the work supported by the grant at the workshop "Big Data & Collaboration" held at Duke University in Durham, NC, USA.

**Conferences:** the last approach pursued to dissemination of our research and findings was the submission to top-tier academic conferences. This was a crucial step of the process, not just for dissemination, but also to engender feedback and better inform the next stages of the research project as it moved forward. Bastos presented his findings at the 48th Hawaii International Conference on System Sciences in Kauai, Hawai; the 7th IEEE/ASE International Conference on Social Computing in Beijing, China; the 2nd IEEE/ASE International Conference on Big Data Science and Computing, also in Beijing, China; the 2014 General Online Research in Cologne, Germany; and the 2014 Digital Humanities Berlin, in Berlin, Germany. In addition to conference proceedings, the results of this project were also featured on the peer-reviewed, open-access online journal PLOS ONE.

* What do you plan to do during the next reporting period to accomplish the goals? (8,000 characters)

Nothing to Report

**Supporting Files**

ANNEX II - HASTAC Scholars Survey Report.pdf: Report of Qualtrics survey that measured the correlation between online collaboration and academic output.

Products Section


Publications (4 items)


Technologies or techniques (1 item)


Inventions, patent applications, and/or licenses (0 items)

Nothing to Report

Websites (1 item)


Supporting Files

ANNEX IV - SocialMediaMineR Documentation.pdf: Documentation of SocialMediaMineR 0.1.
Participants Section

Participants & Other Collaborating Organizations - Who has been involved?

* What individuals have worked on the project?

<table>
<thead>
<tr>
<th>Name</th>
<th>Most Senior Project Role</th>
<th>Nearest Person Month Worked</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cathy N Davidson</td>
<td>Co-founder, HASTAC, PD/PI</td>
<td>.72 person months/115 hours</td>
<td>(August 2012-July 2014, calculated at 5 hours a month)</td>
</tr>
</tbody>
</table>

Cathy N Davidson provided PI support, oversight, and direction to the project. Davidson is an uncompensated PI on this project.

Postdoctoral Fellows

Marco T Bastos, Postdoctoral Fellow, 10 person months, (October 2013 - July 2014, 1.0 FTE)  
David Sparks, Postdoctoral Fellow, 8 person months, (January 2013 - August 2013, 1.0 FTE)

The Postdoctoral Fellows were responsible for conducting all research activities proposed in the grant proposal over the grant lifetime, including conducting social network and data analysis, presenting research findings at appropriate conferences and forums, and completing this final report of research findings to be shared with the academic community. Bastos and Sparks have collaborated in the EAGER Collection and Bastos also collaborated with international scholars during the research that led to the academic papers "Writing styles and knowledge dissemination in academic networks," "What's in a digital humanities blog? A comparison of HASTAC and Hypotheses," and "How Digital are the Digital Humanities? An Analysis of Two Scholarly Platforms" (see section "Key outcomes or Other achievements" for details).

HASTAC Technical Support Staff

Ruby Sinreich, Director of New Media Strategy, HASTAC, 4 person months (October 2012 - February 2014, .25FTE)

Demos Orphanides, Drupal Webmaster & Online Community Strategist, HASTAC .375 person months/60 hours (February 2014- July 2014, .25 FTE)

The HASTAC Tech Support Staff assisted the Postdoctoral Fellows with establishing access to HASTAC.org data, facilitating data access, extraction, and transfer, as well as providing ongoing technical support to proposed grant activities (including setting up and facilitating webcast of invitational workshop and conducting tech support for all grant activities) and Postdoctoral Fellows as needed regarding issues relating to the HASTAC network.
* What other organizations have been involved as partners?*
Nothing to Report

* Have other collaborators or contacts been involved?*

On May 28, 2014 we held the one-day workshop and conference "Big (and messy) Data & Collaboration" at the Duke University PhD Lab on Digital Knowledge to discuss the impact of scholarly networks to cross-disciplinary, multi-institutional research and to inquire into the use of big (and sometimes messy) data in academic, collaborative settings. The event attracted more than 40 participants and the following attendees have collaborated on this project in various capacities:

1. James Moody, Robert O. Keohane Professor of Sociology, Duke University
2. Richard Marciano, Professor, School of Information and Library Science and Affiliated Professor, American Studies, UNC-Chapel Hill; Director, Sustainable Archives & Leveraging Technologies group (SALT)
3. Jonathon Cummings, Associate Professor of Business, Duke University
4. Thomas Nechyba, Professor of Economics and Public Policy and Director of EcoTeach Center, Duke University
5. Kevin Franklin, Executive Director, Institute for Computing in the Humanities, Arts, and Social Science (ICHASS), Research Professor Education Policy, Organization and Leadership, University of Illinois, Urbana-Champaign
6. Vincent Conitzer, Sally Dalton Robinson Professor of Computer Science and Professor of Economics, Duke University
7. Lawrence Carin, William H. Younger Professor of Engineering, Duke University
8. Alan Blatecky, RTI International
10. Angela Zoss, Duke University
11. Barry Peddycord, North Carolina State University
12. Christopher Cline, The Water Institute at UNC
13. Conal Ho, Consortium of Humanities Centers and Institutes
14. Cornelius Puschmann, Humboldt University of Berlin
15. Danica Savonick, CUNY
16. Emanuel Fiano, Duke University
17. Eric Monson, Duke University
18. Fiona Barnett, HASTAC and Duke University
19. Jade Davis, HASTAC and UNC
20. Jenny Korn, University of Illinois at Chicago
21. Joel Herndon, Duke University
22. Kalle Westerling, CUNY
23. Katrin Weller, Leibniz-Institute for the Social Sciences
24. Kaysi Holman, HASTAC
25. Lisa Tagliaferr, CUNY
26. Liz Milewicz, Duke University
27. Lynn Moore, Mozilla Foundation
28. Lynne Steuerle Schofield, Swarthmore
29. Marco Bastos, HASTAC
30. Matthew Gold, CUNY
31. Michelle Ferrier, Ohio University
32. Sheryl Grant, HASTAC
33. Stacy Roberts, North Carolina State University
34. Victoria Szabo, Duke University
35. Will Shaw, Duke University
Impact Section

Impact - What is the impact of the project? How has it contributed?

* What is the impact on the development of the principal discipline(s) of the project? (8,000 characters)

The research supported by the EAGER grant has revealed important relationships in the interplay between online and offline collaboration networks and has shown that the level of engagement in scholarly networks is associated with academic output. Although this relationship suggests that networked young scholars can possibly "outcompete" more traditional peers that refrain from engaging in online scholarly networks, as most scholarship published by HASTAC Scholars in the period was done together with users of the network (77%), we would caution against overemphasizing the impact of social networks on academic output and would rather emphasize the role played by macro-level interactions between university groups and micro-level variables associated with Scholars' personal commitment to academic research. To our knowledge, this is the first attempt to investigate the impact of scholarly social networking on academic output, and we expect the work supported by this grant to inform future research focusing on the interplay between interdisciplinarity and scholarly networks to the academic output of young scholars.

In terms of disciplinary focus, the principal disciplines of the project were Computer Science, Information Science, Education, New Media Studies, Statistics, and the full range of humanities and social science disciplines. In addition to these broad disciplinary categories, several cross-disciplinary areas are the foci of the project: social network analysis, mentorship, digital humanities, and data visualization. We have adapted and extended several social network analysis techniques to accommodate the nature of our data. Specifically, we have developed ways to approach building and drawing network graphs when dealing with bimodal network data in which we observe heavily skewed distributions. For example, the HASTAC.org data makes connections between individual users and the groups on the site to which they belong -- there are several hundred of these groups, most of which have only a few members, but a small number of which have several hundred members. Such "long-tail" phenomena are in some ways difficult to interpret and visualize as a network, since several nodes are typically connected to the majority of other nodes, leading to a very dense network.

Thus, rather than depending exclusively on simple, untransformed adjacency matrices, much of our analysis relies on logged odds ratios of co-occurrence as a measure of association. This lends itself well to dimensionality reduction techniques, allowing for simplified latent spatial representations of network phenomena, including the semantics of tags, group similarities, and member interests. This ability to explore the latent dimensionality of discussion and interests on HASTAC.org is both enabling our study of the digital humanities, as they exist within HASTAC,
and offers a new methodological approach for digital humanists to use. We also implemented several visualization techniques, from GIS to longitudinal social network analysis, which we are the first to apply to the digital humanities data with which we are working. We have chosen the R statistical programming language for most of the programming associated with this research and a portion of this code is now available for other researchers on the CRAN repository.

The impact of the grant also benefited from two lines of academic inquiry pursued during year two. First, the decision to gather data from other scholarly social networks to provide a baseline of comparison to HASTAC.org has contributed to our understanding of online scholarly networks at large and scholarly networks dedicated to interdisciplinary research in particular. Secondly, the deployment and analysis of the HASTAC Scholars Survey allowed for addressing the core questions of the grant by investigating the relationships between publication counts, co-authorship, disciplinary background, and website activity of HASTAC Scholars. These external data sources were used in conjunction with data from HASTAC.org and proved pivotal to understanding how the HASTAC community influences scholarly activity outside of the network.

* What is the impact on other disciplines? (8,000 characters)

The intellectual merits of the project include modeling a method by which cyberinfrastructure can enable successful collaborative research adaptable to other problem-based research centers and scalable to other projects, such as NSF-funded Science of Learning Centers (SLC); determining the best incentives for risk-taking early career interdisciplinary research; and calculating how voluntary, scholarly networking and virtual peer-mentoring support and promote early-stage researchers in both scholarly innovation and high-prestige academic achievement.

The broader impact of the study is in addressing one of the most urgent structural problems in higher education and one on which there is little quantitative, data-based research: how to incentivize and support the range of cross-disciplinary researchers who must work together successfully in order to solve society’s biggest problems. At its broadest, we anticipate that the results of this research will intervene in current theories of organizational behavior, social media protocols, disciplinary and interdisciplinary formation, and self-governing systems, and lay the groundwork for increased collaborations dedicated to tackling the major problems and Grand Challenges of the 21st century.

* What is the impact on the development of human resources? (8,000 characters)

One of our primary areas of inquiry was into networked mentorship, as manifested across HASTAC and particularly through communication on HASTAC.org. HASTAC has been in existence for over ten years, and is bigger and more robust than ever. In light of this, much of our focus has been on learning what structures and patterns have evolved over time to engender this
thriving, interdisciplinary community, and to discern whether these infrastructures (formal or informal) could be replicated elsewhere. In particular, we feel that this study may make an impact on the development of human resources through a better understanding of how mentorship operates when mentor/mentee relationships across geographic, institutional, and even disciplinary boundaries. Although redesigned, this has remained a secondary objective of our HASTAC Scholars survey that provided valuable information to understanding how participation in the online network relates to scholarly productivity and career outcomes. To the extent that we can connect these two spheres, it suggests that the successful HASTAC model can be reproduced in other contexts, including other disciplines, and other interdisciplinary working groups.

* What is the impact on physical resources that form infrastructure? (8,000 characters)

As part of this research project, we released an API for HASTAC.org data with the new iteration of the website. Although the API still requires some adjustments, it should soon become available and prove to be a valuable infrastructure resource for the academic community. As a result, HASTAC data is now available for other researchers interested in this online scholarly network.

* What is the impact on institutional resources that form infrastructure? (8,000 characters)

First of all, the project impacted on HASTAC itself. The HASTAC.org website, on which having an account constitutes becoming a member of HASTAC itself, has gone through considerable redesign in 2014. The data being studied, and the research completed under this grant, informed this redesign process. For example, the latent semantic analysis of tags and the topic modeling of user-contributed content have identified the most prominent subjects of interest on the site (essentially, digital humanities, new pedagogies, and the various DML initiatives). Insight from the HASTAC.org database has also provided exemplar use-cases for the site, around which future functionality can be built.

After extensive data mining, we came to realize, for example, that a relatively small number of users -- typically HASTAC Scholars and DML Competition participants -- are responsible for most of the activity on the site and the redesign is scheduled to take their specific interests into account. However, and considering that a much larger number of individuals come to the site for information, news, and research, the next version of the website also emphasizes the ease of surfacing information. Since the HASTAC.org website is itself the primary venue of communication for HASTAC as an institution, this research-informed redesign helped shape its infrastructure. Our hope is that our findings about the causes and effects of HASTAC’s success as an institution can also be applied to other, similar, interdisciplinary communities.

* What is the impact on information resources that form infrastructure? (8,000 characters)
As part of this research project, we released an API for HASTAC.org data with the new iteration of the website. This is a valuable infrastructure resource for the academic community.

* What is the impact on technology transfer? (8,000 characters)

How to study the impact of virtual associations designed for maximum community innovation is a matter of urgency to many commercial enterprises. SEO, impact, and other ways of studying influence are increasingly contested and our findings, using HASTAC as an exemplar innovative collaboratory, could guide others looking to understand how innovation actually works to shape data (rather than in how conventional data reflect innovation).

* What is the impact on society beyond science and technology?

The urgent social problems facing the world in the current moment -- among them the Grand Challenges identified by the National Science Foundation such as energy, water sustainability, and human sciences and policy design -- require bold, coordinated interventions from academics and policy-makers across specializations. As Myron Gutmann of the NSF has observed, interdisciplinary collaboration is crucial to solving these Grand Challenges. Despite this urgency, university work has tended to take place in disciplinary silos and there is little research that sheds light on what can be done to incentivize and support researchers who move beyond their disciplinary boundaries to work together on large-scale projects. This is particularly true for those projects that require scientists to collaborate with humanities and social sciences scholars focused on the cultural obstacles to the implementation of their scientific and technological solutions.

The development of modern cyberinfrastructure has increased the potential for collaboration among radically disparate geographical and multidisciplinary groups, and the potential for cross-disciplinary and cross-institutional collaboration increases exponentially with each passing year. Yet, to date, scholars have not examined how this network wealth can transform the institutional formations that inhibit work across the disciplinary boundaries and protocols of higher education. To fill this gap, our project mobilized the extensive HASTAC data pool, gathered over 7 years among its user network of approximately 13,000 users, to conduct social networking analysis, data extraction, text and content analysis, and procedural and organizational analysis that provided critically needed evidence for the current and future capacities of institutions to support the interdisciplinary collaborations essential to meet the Grand Challenges confronting the world.
Changes/Problems Section

Notifications and Request

* Changes in approach and reasons for change (8,000 characters)

Nothing to Report. This is the final report of the grant.

* Actual or Anticipated problems or delays and actions or plans to resolve them (8,000 characters)

Nothing to Report. This is the final report of the grant.

* Changes that have significant impact on expenditures (8,000 characters)

There have been two major developments that impacted the budget. Firstly, the hiring of our postdoctoral fellow Marco Bastos on October 2013 had an important impact on the grant expenditures. Dr. Bastos worked on the grant remotely in London, and therefore the project required an increase in the travel budget portion of the grant. This was offset by lower salary costs than what was previously allocated in the grant (which exceeded Duke approved salaries). Secondly, the Conference "Big Data & Collaboration" held at Duke University PhD Lab on Digital Knowledge on May 28, 2014 (with the participation of 40+ attendees) much exceeded the financial provisions previously included in the grant budget. A portion of this additional expenditure was offset by contributions by HASTAC and the PhD Lab.

* Significant changes in use or care of human subjects (8,000 characters)

Due to the considerable redesign of the June 2013 version of the HASTAC Scholars survey questionnaire, we submitted all changes to the research protocol for prospective IRB approval of Duke University's Institutional Review Board. IRB approval was granted on January 3, 2014 and since then the protocol has been renewed (IRB renewal approval was granted on June 11, 2014 with expiration date on June 13, 2015).

* Significant changes in use or care of vertebrate animals (8,000 characters)

Nothing to Report

* Significant changes in use or care of biohazards

Nothing to Report