Massively Distributed Authorship of Academic Papers

Bill Tomlinson (wmt@uci.edu), Joel Ross (jwross@uci.edu), Paul André (paulandre@cmu.edu), Eric P. S. Baumer (ericpsb@cornell.edu), Donald J. Patterson (djp3@ics.uci.edu), Joseph Corneli (holtzermann17@gmail.com), Martin Mahaux (martin.mahaux@fundp.ac.be), Syavash Nobarany (nobarany@cs.ubc.ca), Nithya Sambasivan (nsambasi@uci.edu), Marco Lazzari (marco.lazzari@unibg.it), Birgit Penzenstadler (penzenst@in.tum.de), Andrew W. Torrance (torrance@ku.edu), David J. Callele (dcallele@gmail.com), Gary M. Olson (gary.olson@uci.edu), Six Silberman (six@wtf.tw), Marcus Ständer (staender@tk.informatik.tu-darmstadt.de), Fabio Romancini Palamedi (fabio.palamedi@comtec.pro.br), Albert Ali Salah (salah@boun.edu.tr), Eric Morrill (epmorrill@gmail.com), Xavier Franch (franch@essi.upc.edu), Florian 'Floyd' Mueller (floyd@floydmueller.com), Joseph 'Jofish' Kaye (jofish.kaye@nokia.com), Rebecca W. Black (rwblack@uci.edu), Marisa L. Cohn (mlcohn@ics.uci.edu), Patrick C. Shih (patshih@ics.uci.edu), Johanna Brewer (deadroxy@frestyl.com), Nitesh Goyal (ngoyal@cs.cornell.edu), Pirjo Näkki (pirjo.nakki@vtt.fi), Jeff Huang (wikidemia@jeffhuang.com), Nilufar Baghaei (nilufar.baghaei@gmail.com), Craig Saper (csaper@umbc.edu)

CHI 2012, May 5–10, 2012, Austin, TX, USA.
ACM xxx-x-xxxx-xxxx-x/xx/xx.

Abstract
Wiki-like or crowdsourcing models of collaboration can provide a number of benefits to academic work. These techniques may engage expertise from different disciplines, and potentially increase productivity. This paper presents a model of massively distributed collaborative authorship of academic papers. This model, developed by a collective of 31 authors, identifies key tools and techniques that would be necessary or useful to the writing process. The process of collaboratively writing this paper was used to discover, negotiate, and document issues in massively authored scholarship. Our work provides the first extensive discussion of the experiential aspects of large-scale collaborative research.

Keywords
Collaboration, writing, crowdsourcing, scholarship.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors, Management.

Introduction
The way people collaborate continues to evolve rapidly with technology [21]. Many of the challenges that

researchers face require interdisciplinary input and collaboration [12], including in fields such as HCI. In this paper, we argue that wiki-like and crowdsourced [7] models of collaboration can provide currently-unrealized value to academic writing and research. We propose that such a model of engaging many people from different disciplines through short-time-scale cycles of collaboration can enable research and provide insights that go beyond what is feasible with the small-group model that is prevalent today.

Many massively multi-authored academic papers already exist, both as targeted experiments, such as the Polymath Project [9], and as the result of large-scale collaborations in particular disciplines, such as Venter et al.'s "The Sequence of the Human Genome" [23] (with contributions from approximately 2900 co-authors). However, these papers do not discuss the experiential aspects of such collaborations.

The current paper was written in a distributed fashion by 31 authors. At the onset the only aspects of the paper that were specified were the topic of the paper, the target venue (with format and deadline constraints), and a few principles about collaboration. Writing the paper provided the research context, and the final paper emerged as a research artifact. We have sought to engage in a conscious reflection on the authorial process, and to identify possibilities and challenges that arise at the juncture of academic research, distributed authorship, and digital technology.

The goals of this paper are twofold. The first goal is to discuss the methodology and tools that enable an *ad hoc* group of academic researchers and other interested individuals to gather, work together and produce a scientific paper collaboratively. Working together with a large number of authors, we must take into account the preferences and ideas of a broad cross-section of researchers, a process known to be challenging [10, 19]. The second goal is to experiment with this type of collaboration, including empirically finding and validating best practices, and reflecting on them.

Our work provides the first extensive discussion of the *experiential* aspects of large-scale collaborative research. Our findings were established via a research-through-writing approach in line with earlier discussions of research-through-design [26]. This work is of value to CHI practitioners seeking to design tools and environments for large-scale collaboration on scientific research and writing, by revealing the opportunities and complexities of this process.

Related work

Collaborative writing has been around since well before digital computation, but contemporary projects such as Wikipedia and Crowdforge [13], have demonstrated that crowdsourcing could be used to create written content of great scope and high quality.

Work on the social aspects of *coordination in collaborative writing* in the Wikipedia context [24] particularly informed the analysis we give in this paper, but we see many differences between writing for a general audience without deadlines, and writing a peer-reviewed submission with deadlines. These differences suggest different interaction designs. The closest empirical paper of which we are aware examined the use of a collaborative online tool [3], but for an artificial writing task instead of a research publication. By contrast, we found it most convenient to gather data on

collaborative writing in a research context by devising a real research situation in which we could make observations.

Coordination Mechanisms
In current academic practice, significant grants are often awarded to multi-institution and multi-discipline groups. In these projects, which can span years, there are often group meetings or phone calls to set an initial direction, with each institution taking a specific aspect of the work and developing it largely independently. Small teams can work more closely, but this takes significant effort (in terms of availability, direction, and roles). These examples suggest different models of coordination, as Van de Ven and others [22] have distinguished: pooled, sequential, reciprocal, and team modes.

With complex work such as authoring an academic paper, coordination is key to producing coherent and high-quality output. Different circumstances may have different coordination mechanisms that are suitable, for example: markets (first come first served), hierarchical management (project leaders), standardization (pre-defined templates), communication (discussion pages and talk between authors), and shared mental models (initial structure provided by type of community or template) [6, 14, 15].

Congruent Methodologies
Our approach to a massively multiauthored paper borrows from, but also differs from, the "talk aloud method" and "protocol analysis" [8]. The "bias" in the way we conducted our surveys is inherent in research-through-writing. In the future, multi-authors could build on our approach, for example through an auto-ethnographic methodology [16].

Process

To provide a scaffold for the writing of this paper, a draft of an abstract and a few basic collaboration policies were established by the initial corresponding author (ICA) of the project, and specified in a non-editable web page: http://www.ics.uci.edu/~wmt/CHI2012CollaborativePaper.html.

The initial goal was to write a CHI 2012 full paper. After asking several colleagues for feedback on the core principles, the ICA began recruiting other authors.

During the first several weeks of the writing, the authors collaborated using a Google Doc. However, at the time of the influx of new authors from the CHI mailing list, the Google Doc encountered an apparently unfixable error, perhaps having to do with the large number of authors (>25) attached to the project. The production was therefore moved to Etherpad, specifically, to the public installation hosted on PiratePad.net. EtherPad allowed the authors to finish a complete draft of the paper. Unfortunately, PiratePad.net encountered an error that prevented the team from viewing the history of the document.

In the final weeks, an author survey was conducted and integrated into the paper. The paper was formatted and shortened to conform to the page limit, with abundant additional written content, notes, references, etc. remaining available in the online versions of the paper: https://docs.google.com/document/d/1epSu3hbGp0eafDHiqRLeWqezrRcOEdgi44KLa4TiH40/edit

http://piratepad.net/Massively-Distributed-Authorship-of-Academic-Papers

The final edits were done via an online, shared, and synchronized folder (Dropbox), in which subsequent rapid revisions were created as separate files. With one week remaining, the ICA analyzed the paper with the TurnItIn plagiarism detector, to confirm that the paper did not contain any accidental or malicious instances of plagiarism.

While the CHI full paper was not accepted, it was strongly recommended for alt.chi. Based on the feedback in the reviews, the collective of authors used PiratePad and then MS-Word to revise the text and reformat it for alt.chi.

Qualitative evaluation of participation

We conducted a qualitative evaluation of the project to understand authors' experiences of writing and contributing to a massively distributed academic paper. We used surveys with open-ended, multi-faceted questions centered around a specific theme. For example, the thematic prompt, "What did you think of the process of writing the paper?" suggested such questions as "What worked well? What didn't? Was there anything confusing? What kinds of additional tools might help? Do you think this process would adapt well to writing other papers?"

The surveys themselves were designed collaboratively. A portion of the EtherPad document was marked as dedicated to the qualitative evaluation, and numerous authors added questions that might be included as part of a qualitative interview. Some authors also added pointers, suggesting links between evaluation questions and other parts of the paper. Two authors then edited the list of questions to reduce redundancy and improve coherence; this edited question list was then used to create a survey that was sent to all authors. The survey responses were collected over a five-day period that ended three weeks before the submission deadline. This schedule was chosen to allow sufficient time for analyzing and writing up the results of the surveys. The results of the evaluation were included in the draft sent to all authors for editing, approximately two weeks before the deadline.

Participation and Interaction

Most authors described their participation style as "editing and commenting." In the words of one author, "participation style has been mostly helicopter-commenting."

Some problems arose due to the "recursive" nature of this paper, which confused some co-authors (and reviewers). The writing process resembled a game of Nomic [20], where the game is about changing the rules of the game itself. For example, arguments in one section of the paper might use quotes from another section as evidence. One author suggested that the outcome did a "poor job [of] collectively replicating Malcolm Ashmore's Reflective Thesis" [2]. The self-reflective nature of the papers led one of the CHI reviewers to describe the paper as "Seinfeldish".

Over time, most authors followed one of two participation trajectories. Both trajectories started with initial excitement, including either reading through the then-current draft of the paper, leaving comments in some sections, responding to other authors' comments, writing somewhere between a couple sentences and an

entire section, adding their name to the author list, or some combination thereof. In the first trajectory, this initial bout of activity was followed by a slow, or in some cases rapid, decline, with some authors dropping off the paper entirely.

The second trajectory, taken by fewer participants, pushed the paper through submission, and revision. In response to one of our survey questions, all authors whose names appear on the final paper unanimously responded that they would like to participate in distributed, collaborative online writing in the future.

Distributed Authorship, Distributed Authority
Along with the relatively low levels of parallel communication about the paper (most discussion happened in the paper itself, with limited backchannel discussions), authors also felt a lack of any centralized control directing the paper. Contrasting it with other multi-author efforts in which he was involved, one author said that "with this paper, no one's in charge." Not having a centralized voice of authority become problematic in that the central thesis of the paper was not always clear.

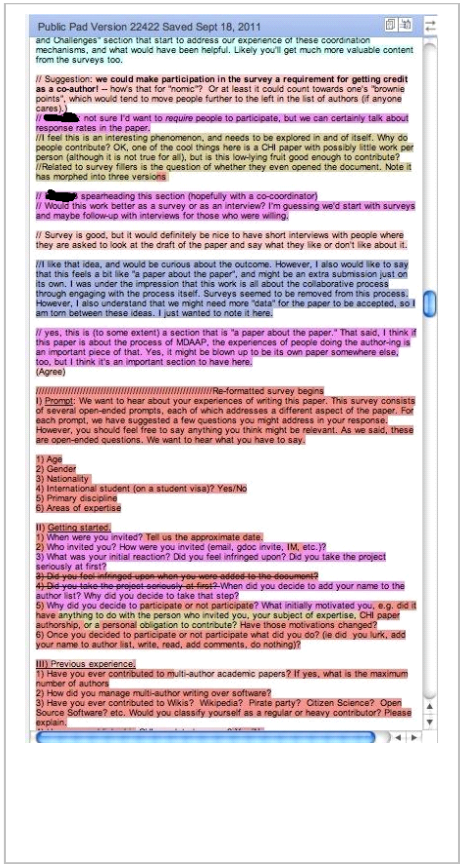
Some appreciated this: one author enthusiastically compared the writing process to participation in a seminar. Others pointed to the concerns about editorial authority associated with removing content and finding consensus in multiple voices. In the words of one author, "I think it is quite chaotic to find a common structure and it is sometimes unclear: 3 authors comment and who decides finally? The main author?"

These responses suggest an implicit assertion that it is necessary for the author(s) working on each section to understand how that section fits into the larger structure of the paper. The intermittent sense of confusion about the paper's aims reflects a certain assumption that many the authors on this paper have about the academic writing process: that in order to be effective, all authors must have a clear understanding of the larger argument to which they are contributing.

Medium of Collaboration
In writing this paper, the authors discussed a number of potential tools —existing and hypothetical— that could support the distributed authorship of academic papers. The most common type of tool considered were systems that support (simultaneous) multi-editing (cf. [17]). Some authors also suggested using a system like Scribtex (scribtex.com), though this would require working knowledge of LaTeX and so could have added a barrier to entry for participating in the collaboration. Other systems such as the CoWord plug-in for Microsoft Word could also have enabled simultaneous writing and editing.

However, it is possible that for the number of authors on the scale of this particular paper, viewing live simultaneous edits may be less important than simply having on-demand access and being able to track contributed text (who wrote what). For this, version control systems, such as Assembla (assembla.com) or Git might have been more effective.

Other tools can be integrated into the writing process to support particular sub-tasks. For example, the authors used Zotero (zotero.org) to collaboratively build a reference library: https://www.zotero.org/groups/mdaap.



Although online tools appear to be a natural solution to the problem of working across time zones, they are, in their current state, lacking in support for multiple authors and voices. Authors supplemented the messiness of multi-authorship through other tools, such as EtherPad's built-in chat and e-mail. One author told us, "The chat was really nice—it was better than commenting in the paper directly since it didn't add additional 'noise'."

The technology, together with our governance choices, led to an emergent text that only weakly permits tracing of individual contributions. It was not obvious if a particular piece of text came from a graduate student, a researcher in industry, or a professor (which could be seen as a good thing). A screenshot - Figure 1 at left, available at higher resolution at: http://postimage.org/image/ym6n060cx) - illustrates how things looked to collaborators during our writing process.

Challenges of Integrating with Existing Practices and Publishing Infrastructures

Existing practices and infrastructures for the scientific publication process in our discipline are designed with relatively low numbers of co-authors in mind, and generally do not effectively support this kind of writing. For example:

ONLINE SUBMISSION SYSTEMS

Manually entering author data for a paper with thousands, hundreds, or even tens of authors can be a tedious process (if managed by one author), as supported by the current submission systems.

PEER-REVIEW PRACTICES

Many journals and conferences require anonymized submissions; however this type of article is difficult to make anonymous because a large number of researchers have been aware of the research effort (e.g. the author recruitment process involves broadcasting to a large community). Reviewer assignment is also challenging; many qualified reviewers may have been involved in the process or may have conflicts of interest with at least one of the co-authors. This issue can become problematic for program committee meetings as well, where members of the committee who have a conflict of interest in evaluating a paper must not participate in the discussion of the paper.

IMPACT MEASUREMENT PRACTICES

Some current practices for assessing the productivity and impact of researchers do not consider their relative contributions to papers. Two standard indices for measuring the impact of researchers, H-Index and G-Index, do not take into account the differences in the levels of contributions of co-authors. Thus, the growth of this form of scholarship may make these indices less useful as an estimate for academic impact.

The author list is not blank, nor replaced with a collective pseudonym (like *Bourbaki or D.H.J Polymath*), because the authors wished to acknowledge and take credit for their contributions, and agreed to be scientifically accountable for the content as a whole. We agreed that the author list should be ordered by placement "bids" from authors, with the ICA's judgment used to resolve ties.

DIGITAL LIBRARIES AND PUBLICATION FORMAT

Reference formats are also designed for small numbers of co-authors. For example, the APA and the Vancouver convention [11] limit the number of authors. This can demotivate authors who will not be among the top listed authors of a paper. Large author lists take up valuable space in the original publication, but will almost certainly not appear in future citations due to the annotation effort and/or space required. Moreover, even when browsing through papers in digital libraries or indexing systems that provide full metadata of papers, it may be of little benefit to browse through a long list of authors, not knowing how each of them has contributed to the paper.

Discussion

We will now take a look at various strategies that can be used to mitigate the risks associated with our approach. We begin by summarizing the main challenges.

1. More authors means more content, but also more words thrown away. Many of the words written by authors were deleted during the ongoing editing process. The sheer mass of deleted words might raise the question of whether authoring a paper in such a massively distributed fashion is efficient.

2. Technology provides inadequate support for distributed authoring. During the time we have been working on this project, we have tried, discovered, or created various tools for collaboration, but none of them appeared to be a silver bullet for all our needs.

3. Task and domain differences. There is a distinction between "collaboration on research" and "collaboration on writing". Although one normally follows the other, certain models of widespread collaboration might focus on, say, outsourcing experimentation (e.g., ScienceExchange.com), or ask for expert hands/eyes on certain parts of a paper (e.g., literature review, data analysis).

Strategy 1: Know Thy Collaborator (and Granular Contributions)

This project had many contributors, most of whom did not know one another. Additionally, several were from outside of HCI, and at least one from outside academia. The nature of the project meant that most participants added miscellaneous comments or perhaps took ownership of a section. At the end of the process, a small set of authors did significant restructuring and rewriting. Traditional collaborators often know and trust one another, with clearly designated leaders or first authors. Future collaborations should further explore the role of management (and associated functions such as task decomposition and assignment), relative to measures such as quality, coverage, coherence, and creativity.

Strategy 2: Improved Tool Design

From our experience it is clear that a system for writing massively distributed academic papers needs some way of maintaining and explicating the current state of the paper. At a high level, incoming authors need to understand what "phase" the paper is in (e.g., planning, writing, editing). Future tool design might look to decision support system literature [18] for mechanisms to aid the collaborative process. Sections that are "complete" may also need to be safeguarded, and potential editors pointed towards areas that still need further work. Having a strong sense of structure

could better enable such collaboration—although this paper began with a rough (section-level) outline, the structure became harder to follow as more sections were added and comments were written in-line. We attempted to maintain an up-to-date high-level outline of the paper as time went by, but this process could have been automated by a more sophisticated tool. More broadly, there is currently no simple way to curate or challenge a portion of the text, elicit responses from the authors, and apply changes that agree with a majority opinion or the opinion of more senior contributors that the authoring community might wish to empower.

Strategy 3: Develop Suitable Coordination Mechanisms
In writing this paper, it seems that communication between co-authors happened mostly in the context of the document itself. However, we also found that when authors come and go in the online environment, previous discussions are often brought up that had already been resolved, making the process repetitive at times. It has been a common practice in collaborative authoring tools to provide various forms of annotations to support the coordination of the authoring process [4]. Here, the primary method of coordinating activities was free-form annotations in the document. No annotation schema or rules were defined in advance, and the participants used various methods for distinguishing notes from content, such as using font styles (italic, bold, all-capitals), programming conventions (e.g., starting with //) or changing font color. For higher-level coordination needs, towards the end of the writing process, the ICA began using email to send out alerts and keep writers-up-to-date on process. This switch from the initial "hands-off" approach provided the structure we needed in order to meet deadlines. Further thought about how to best shape and signpost the project (both from a "writing" and a "research" standpoint) would help in future efforts.

Future work

We have sought to identify challenges and opportunities in massively distributed authorship of academic papers, as well as reflections on the process and considerations for future researchers. Further experiments and repetition are required both to validate the analysis suggested here and to explore the relationship between this form of collaboration and different forms of research. For example, what research disciplines are best supported by this method of collaboration? What forms of research (e.g., user studies, design visions)? Future efforts along these lines might try some alternate methodologies (e.g., discourse analysis [25]).

Future work is also needed to develop the tools that can enable massively distributed authorship. For example, a website to act as a clearing-house for in-progress papers, ideas, or initial findings, would give authors a central location in which to contribute, as well as find new collaborators or stumble across relevant ideas, potentially aiding serendipity, insight, and discovery (as in [1]). Building an authorship community interested in shared process and/or content may help address the indifferent feelings many authors experienced about participation in this experimental effort.

Finally, this form of collaboration could also be extended to other parts of the academic publication process. For example, additional features could be

integrated with WikiCFP (wikicfp.com) so that authors can collectively determine what papers are being written for a given conference. Conferences and workshops could also produce post-event position papers written by the participants collectively, or include it in their proceedings a collaboratively-written summary of the event.

Conclusion

Massively collaborative crowdsourced projects, such as Wikipedia and Linux, affect aspects of many people's everyday lives. Open Data is a related emerging practice in the hard and computational sciences. The corresponding practices for "human sciences" are at present somewhat less clear.

In this paper, we consider one aspect of massively collaborative research that has potential to alter the way we work and discover: authoring. We identify challenges in: participation models for crowdsourced authorship; tool design and coordination mechanisms for supporting stages of a research and authoring process; and piecemeal contribution versus ownership and engagement. There are philosophical and practical questions with regard to contribution and authorship that we have only begun to discuss.

Academia itself is a massively collaborative undertaking, but massively multi-author writing of the kind described in this paper is relatively new. Nevertheless, with the spread of communication tools that facilitate various aspects of the process, and the exploration of new techniques for working together on large scales, this form of scholarship may begin to play an increasingly salient role in the pursuit of knowledge. When deployed well, collaboration can help our community become even more effective in its efforts to "contribute to society and human well-being."
[5] Research funds should be targeted towards developing more effective and rapid means to solve a range of research problems. Without sufficient investment in this domain, we will be bypassing some of the great advantages of living in a networked world.

Acknowledgments

Many more people contributed to this paper than those currently on the author list. At times, the paper had as many as 38 people on the author list, and many others read drafts and propagated information about the project. The CHI reviewers provided helpful insights as well, and some of their ideas are in this version of the paper (although since they were anonymous, we were not able to invite them to be on the author list.) The authors thank the numerous institutions and other entities that have provided support for their respective participation in this research.

References

[1] André, P., schraefel, m. c., Teevan, J. and Dumais, S. T. Discovery Is Never By Chance: Designing for (Un)Serendipity. *Proc. Creativity & Cognition 2009*.

[2] Ashmore, M. *The reflexive thesis: Wrighting sociology of scientific knowledge*. University of Chicago Press, Chicago, IL, USA, 1989.

[3] Brodahl, C., S. Hadjerrouit, and N. K Hansen. 2011. "Collaborative Writing with Web 2.0 Technologies: Education Students." *J. Information Technology Education*: 31.

[4] Cadiz, J.J., Gupta, A. and Grudin, J. Using Web annotations for asynchronous collaboration around documents. *Proc. CSCW 2000* , 309-318.

[5] Code of Ethics — Association for Computing Machinery. http://www.acm.org/about/code-of-ethics/#sect1.

[6] Crowston, K. A coordination theory approach to organizational process design. *Organization Science*. (1997), 157-175.

[7] Doan, A., Ramakrishnan, R. and Halevy, A.Y. Crowdsourcing systems on the world-wide web. *CACM*. 54, 4 (2011), 86-96.

[8] Ericsson, K. A. 2002. "Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis." *Applied Cognitive Psychology* 16 (8): 981-987.

[9] Gowers, T. and Nielsen, M. Massively collaborative mathematics. *Nature*. 461, 7266 (2009), 879-881.

[10] Hinds, P.J. and Bailey, D.E. Out of sight, out of sync: Understanding conflict in distributed teams. *Organization science*. (2003), 615-632.

[11] ICMJE: Uniform Requirements for Manuscripts Submitted to Biomedical Journals. http://www.icmje.org/.

[12] Kim, S. Interdisciplinary cooperation. *Human-computer interaction* (1995), 304-311.

[13] Kittur, A., Smus, B. and Kraut, R. 2011. CrowdForge: crowdsourcing complex work. *Proc. CHI 2011*, 1801-1806.

[14] Malone, T.W. and Crowston, K. What is coordination theory and how can it help design cooperative work systems? *Proc. CSCW 1990*. 357-370.

[15] Malone, T.W., Crowston, K., Lee, J., Pentland, B., Dellarocas, C., Wyner, G., Quimby, J., Osborn, C.S., Bernstein, A. and Herman, G. Tools for inventing organizations: Toward a handbook of organizational processes. *Management Science*. (1999), 425-443.

[16] Muncey, T. 2010. *Creating autoethnographies*. Los Angeles: SAGE.

[17] Olson, J.S., Olson, G.M., Storrøsten, M. and Carter, M. Groupwork close up: a comparison of the group design process with and without a simple group editor. *ACM Trans. Info. Sys.* 11 (1993), 321-348.

[18] Shim, J.P., et al. Past, present, and future of decision support technology, *Decision Support Systems*, 33, 2, (2002), 111-126.

[19] Steiner, I.D. *Group Process and Productivity*. Academic Press, New York, NY, USA, 1972.

[20] Suber, P. *Nomic: A Game of Self-Amendment*. Peter Lang Publishing, New York, NY, USA, 1990.

[21] Tapscott, D. and Williams, A.D. *Wikinomics: How mass collaboration changes everything*. Portfolio Trade, 2008.

[22] Van de Ven, A.H., Delbecq, A.L. and Koenig Jr, R. Determinants of coordination modes within organizations. *Am. Soc. Rev.* (1976), 322-338.

[23] Venter, J.C., et al. The sequence of the human genome. *Science*. 291, 5507 (2001), 1304.

[24] Viegas, F. B, M. Wattenberg, J. Kriss, and F. Van Ham. 2007. Talk before you type: Coordination in Wikipedia. *HICSS 2007.* 78-78. IEEE.

[25] Wohlwend, K.E., Vander Zanden, S., Husbye, N.E. and Kuby, C.R. Navigating discourses in place in the world of Webkinz. *J. of Early Childhood Literacy*. 11, 2 (2011), 141.

[26] Zimmerman, J., J. Forlizzi, and S. Evenson. 2007. Research through design as a method for interaction design research in HCI. CHI 2007, 493-502. ACM.