Online Question Answering Practices to Support Healthcare Data Re-use

Blake, Catherine¹,², Souden, Maria¹, Anderson, Caryn L.¹,², Twidale, Michael¹,², Stelmack, Jenifer E.¹

1. US Department of Veterans Affairs
   VA Information Resource Center
   Edward Hines Jr. VA Hospital
   5000 S. Fifth Ave, Bldg 18 (151V)
   Hines, IL 60141
   Maria.Souden@va.gov, Jenifer.Stelmack@va.gov

2. Graduate School of Library and Information Science
   University of Illinois at Urbana-Champaign
   501 E. Daniel Street, Champaign, IL 61801
   clblake@illinois.edu, clndrsm2@illinois.edu, twidale@illinois.edu

ABSTRACT
Institutional data collection practices inevitably evolve over time, especially in a distributed clinical setting. Clinical and administrative data can improve health and healthcare, but only if researchers ensure that the data is well-aligned to their reuse goals and that they have adequately accounted for changes in data collection practices over time. Our goal is to understand information behaviors of health services data users as they bridge the gap between the historical data and their intended data reuse goals. This project leverages more than a decade of listserv posts related to the use of clinical and administrative data by US Department of Veterans Affairs (VA) employees, providing longitudinal insight into data reuse practices in both research and operational settings. In this paper we report the results of a pilot study that highlighted questions raised in the use of data and the knowledge engaged to answer them.

Keywords
Forums, Big Data, Health, Communities of Practice, Social Question Answering

INTRODUCTION
Organizations have been leveraging data assets long before the term 'big data' popularized this activity. Fifteen years ago researchers noted the enormous potential of using clinical and administrative data collected by the US Department of Veterans Affairs (VA), when they stated that “Access to large populations with centralized record keeping systems greatly facilitates the conduct of epidemiologic research, because of reductions in the costs of performing such research and the higher chances of an adequate sample size to test study hypotheses” (Boyko, Koepsell, Gaziano, Horner, & Feussner, 2000, p307).

The VA maintains a substantial number of datasets for use by health services and operations researchers that are sourced from clinical and electronic health records (EHR), claims, administrative, and operational data. The organization provides these data for use in improving patient outcomes, clinical care, and for organizational processes. This work often involves novel uses and integration of data beyond their original intent. Such data reuse is always a challenge, requiring a deep understanding of the meaning of particular fields and background on how values were collected, coded, and represented. This is particularly challenging if those reusing the data were not involved in the initial collection. Additionally, historic datasets that extend over many years can be extremely valuable for longitudinal studies, but require an understanding of how data collection and clinical care practices have changed over the life of the dataset in order to be able to derive sensible temporal analyses.

The VA Information Resource Center (VIReC) is a VA Health Services Research & Development-funded resource center that provides data knowledge, data documentation, education, and help desk services as well as HSRData-L, a VA listserv for data users. HSRData-L is an email-based virtual community of VA researchers who share their collective knowledge and experience about VA data and information systems. HSRData-L is open to anyone with a VA email address. The 1,130 participants include VA researchers, operations analysts, data stewards, clinicians, policymakers, program office staff, and technical experts.

In examining the HSRData archive, we have a unique opportunity to explore how a community of practice (investigators and analysts using VA data) makes sense of clinical and administrative data in order to address research and quality improvement needs. Since we are analyzing listserv posts over a large time frame (2000-2015), the patterns of behavior that emerge will be less sensitive to particular technologies or users and thus better reflect “typical” data re-use practices.
Although emphasis has been placed on the computational resources needed to process big data, much of the effort needed to leverage longitudinal data assets goes into pre-processing and contextualizing the data. For example, in a healthcare setting analysts must be sure that they have accurately differentiated between a patient episode, encounter, and visit as any misalignment could skew subsequent analysis. Similarly, they must ensure that they have accounted for changes in data collection and clinical care that inevitably evolve over time. For example, a posting from 2004 in the dataset that we are exploring stated that “There are data back to 1992, but the definition of an outpatient visit has changed, and the older count is not comparable to more recent years.”

The inadequacy of an entirely externalized repository approach has been realized by other researchers such as Ackerman et al. who stated that “The information in a repository is easily transferable and reusable, but decontextualized information is often not easy to use. Users often need to find either knowledgeable people or people who can help them apply the information to the current situation or problem” (Ackerman, Pipek, & Wulf, 2003,xii). In the context of big data this implies that collecting and making data available is a necessary but not sufficient step to ensure that data assets are easily and appropriately reused. Additional support is needed to understand the evolving meanings of values in long-term datasets.

Research studies of question asking and answering have a long history and come from a number of perspectives. (Ackerman, Dachtera, Pipek, & Wulf, 2013) give a useful perspective of how Computer Supported Cooperative Work has analyzed the sharing of knowledge and expertise over 20 years and in a variety of settings. Their historical survey reminds us that email distribution lists, online forums, bulletin boards, and Usenet have been used to enable people to ask and answer questions since the early days of the Internet. More recently, online forums have become popular. Various terms have been used to describe interaction in such forums, including, recently, “Social Q&A” (SQA). (Gazan, 2011) overviews some of the communities and the research issues that have been explored. Particular attention has been focused on the quality of the answers provided and the reasons why people bother to give these answers. Unlike our context, most of the studies in Gazan’s review are of public forums where anyone can ask and respond. Despite (or perhaps because of) this openness and low barrier to participation, numerous studies have found that SQA sites provide reasonably high quality answers.

**METHOD**

Grounded theory (Glaser & Strauss, 1967) has often been used to explore forum posts (see (Vaat & Walsham, 2013) for a discussion of such strategies). This study utilized five team members to conduct an iterative, ground-up annotation of forum posts. Three members had direct experience with grounded theory and two had not used grounded theory before, but came to the project with a wealth of experience from related domains. The variations in experience and knowledge domains of the annotators were mitigated through a paired review process in which annotations were reconciled between team members with complementary experience.

Data for the project comprises 184 plain text files of listserv messages, one for each month from January 2000 through April 2015. The text files first were transformed using Java to remove header details and partition into sentences. Quoted text was also marked. Compressed messages, attachments, and duplicate HTML content were excluded. The analysis reported in this paper considers a subset of the overall project data and spans 7 of the 15.5 years available, from 2000-2005 and 2013 (a recent year was included to ensure that the pre-processing methods were robust with respect to changes in the listserv technology). The subset comprises 3,323 posts with nearly 100,000 sentences.

A random sample of 50 forum posts was extracted from the 7-year subset and one of the project members began to annotate each sentence in the first few posts to provide an overview of the annotation process. The study team worked through the annotations together until the group was starting to reach consensus and then each member completed the remaining annotations individually. Individual annotations were then compared and differences discussed which led to a set of high level (level 1) and lower level categories (see results section for a complete set of categories). A second set of 50 posts were distributed to each group member who again annotated each sentence based on the earlier discussions. Then the group met to discuss and resolve differences and a third set of listserv emails were again pulled at random. At this point very few new level 1 and level 2 categories emerged and we transitioned to the second phase of annotation.

In the second phase each group member was assigned as a primary annotator for a set of 50 randomly drawn listserv posts and asked to annotate sentences using the categories identified in the first phase. Once the annotations were complete, they were reviewed by an assigned secondary annotator. Annotating text is a monotonous and tedious activity; many of the differences in the second annotation phase were due to missing annotations rather than a difference of opinion in what the annotation should be. Disagreements between the primary and secondary reviewers were discussed and resolved. Then the entire group met to discuss areas of uncertainty and establish if any new level 1 or level 2 categories should be used in the next round of annotations. Once the first 250 posts were in place a second set of 50 forum posts, again drawn at random, were manually annotated by each of the 5 authors using the same process for a total of 500 posts.
RESULTS AND DISCUSSION

The messages comprised 41% of headers, which were set aside, and the remaining 59% of sentences were annotated.

Characterizing thread patterns

An average of 20 posts per month were made in 2000, which increased to 84 posts per month in 2013. Thus the average number of posts is much less than the number of messages posted daily for support groups (74) or hobby groups (99) (Galegher, Sproull, & Kiesler, 1998), however, the user population for those public forums was much larger than the highly focused 1,130 users who have access to this forum, which is housed behind the VA firewall.

Figure 1 shows the number of posts per thread. There were 269 posts with no responses (i.e., a thread length equal to 1), but closer inspection showed that many were ‘subscribe’ or ‘unsubscribe’ messages, automated ‘out of office’ responses, and calls for conference and events, all of which would not need a public reply. Generally, shorter threads tended to address more straightforward questions, including where to find a particular variable, dataset, or resource. More complex questions concerning data utility and measurement constructs engendered longer discussions, and questions with significant variability or which raised larger issues garnered the most threads.

The average number of responses per post was 3.7 which is more than the 1.8 posts reported in early studies of news forums, (Whittaker, Terveen, Hill, & Cherny, 1998), but consistent with the 3.3 posts reported in a similar listserv where users asked questions and provided answers (Hansen, 2009) and similar to the mean number of 3.5 reported for support groups and 3.9 for hobby groups (Galegher, Sproull, & Kiesler, 1998).

Characterizing sentence content

Initial categories evolved from the 500 randomly selected posts during the annotation process, resulting in six high level categories (see Figure 2). The categories were not mutually exclusive, i.e., a sentence could be assigned to more than one category. The high level categories were Announcements (e.g., invitations to informal question answering events), Administration (e.g., subscribe and unsubscribe requests) and Ignore (ignorable text like auto-respand and contact information), which left us with text that could be coded as Question, Answer, or Discussion. When sentences indicated that the question or inquiry was resolved, that was also captured. However, only 19 sentences clearly stated that the question was solved, so these annotations were collapsed into the more general level 1 category of Answer.

During the iterative annotation process an additional set of level 2 categories evolved to mark phenomena of interest for later unpacking (see Figure 3). Most sentences marked Ignore at level 1 contained contact information or a compressed file name, categories not shown in Figure 3. In level 2 categories, Context was annotated for both questions and answers where users elaborated on the steps they had already tried or cautioned about how to re-contextualize data sources. Level 2 categories also showed that users posted code and links to justify how they found the answer (Source) or directed others to where the answer could be found (Referral). Community norms such as how the users posed a question or addressed the reader were also annotated (as Politeness).

In some cases it was clear that there was a gap in the information provided in a question or answer (coded as Gap). A small number of sentences were annotated as Coming Soon, where future resources were referenced, Agreement, where a request by one user was reiterated by another, and Offlist, where it was clear that the discussion would be continued by email. This last finding is consistent with earlier work showing that 40% of users received answers directly by email (Lakhani & von Hippel, 2003).
CONCLUSIONS AND NEXT STEPS
We have described a pilot study of a listserv designed to support a user community which is actively reusing clinical and administrative health data. After collectively developing level 1 and 2 categories to capture the main focus of a sentence, a random subset of 500 listserv messages comprising 12,546 sentences was manually annotated. Although the random selection process provided a good way to review a large cross section of posts, our next round of analysis will sample entire threads so that the gaps in either a question or answer can be more easily identified. Further analysis will characterize the type of questions and answers provided.

With that said, we have learned that the average number of posts per thread in the VA listserv mirror other question-answering listservs. The high-level annotations show that most sentences are part of a questioning or answering activity and the sub-level categories provide insight into how users elaborate on and support these, providing context, referrals, code, and identifying source material.

Ironically, many of the data reuse challenges that users were discussing in their posts were part of our own experience. We overcame challenges in pre-processing the data: making decisions on whether and how to include or exclude types of data and how to adjust the methods for cleaning the data to address inconsistencies in data quality, format, and structure related to diversity of time frames and data provenance. Related and new challenges emerged as we began to categorize our data and organize it for the next phases of analysis, especially in identifying the essence of a post based on small contextual clues and sometimes in absence of the complete thread. These experiences seemed to parallel many of the kinds of experiences we were reading about in the user community emails.

ACKNOWLEDGEMENT/DISCLAIMER
This material is based upon work supported by the Department of Veterans Affairs (VA), Veterans Health Administration, Office of Research and Development, Health Services Research and Development Service (HSR&D). VIReC is funded by HSR&D Grant SDR 98-004. The views expressed in this poster are those of the authors and do not necessarily reflect the position or policy of the VA or the United States government.

REFERENCES