Arkose: Reusing Informal Information from Online Discussions

Kevin K. Nam
School of Information, University of Michigan
1075 Beal Ave., Ann Arbor, MI 48109-2112
ksnam@umich.edu

Mark S. Ackerman
Dept of Computer Science and Engineering, and
School of Information, University of Michigan
1075 Beal Ave., Ann Arbor, MI 48109-2112
ackerm@umich.edu

ABSTRACT
Online discussions such as a large-scale community brainstorming often end up with an unorganized bramble of ideas and topics that are difficult to reuse. A process of *distillation* is needed to boil down a large information space into information that is concise and organized. We take a system-augmented approach to the problem by creating a set of tools with which human editors can collaboratively distill a large amount of informal information.

Two design principles, which we will define as incremental diagenesis and incremental summarization, help editors flexibly distill the informal information. Our system, Arkose, is built as a demonstration of these principles, providing the necessary tools for distillation. These tools include a number of visualization and information retrieval mechanisms, as well as an authoring tool and a navigator for the information space. They support a gradual increase in the order and reusability of the information space and allow various levels of intermediate states of a distillation.

Categories & Subject Descriptors
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Design

Keywords
Information reuse, community knowledge, online communities, knowledge communities, collaborative distillation, information organization, incremental formalization, design rationale, CSCW

1. INTRODUCTION
Communities know a great deal. It's clear that groups of people, especially large or Internet-scale groups, have a greater understanding of a problem and its issues than any individual or even a select committee. We would like to find a way to garner and then reuse that knowledge.

Indeed, as access to the Internet becomes more ubiquitous and the infrastructure for publishing and discussing people's ideas proliferates, it has become common to hold a large group discussion online. For example, this could be used for a brainstorming on product ideas or discussion on new technology deployed within a corporate setting. Governments, institutions, and universities could discuss various future plans for organizational changes in order to reach a "shared mind". Later they might want to revisit or reuse that understanding.

However, a standard problem with online discussions is that once use has ceased either by deadline or by neglect, a site is often a bramble of ideas and topics, too large and unwieldy for its information to be successfully reused. A process of filtering, structuring and organizing of the information, or the process of *distillation* as we call it, is needed [Ackerman and McDonald 1996; Ackerman et al. 2003].

In this paper, we discuss our system augmented approach to distillation with Arkose, a software system we have developed to provide a set of augmentative tools to facilitate the filtering, structuring, and organizing. Its visualizations allow editors to quickly understand the discussion space, as well as function as a substrate for gradually transforming a bramble of nodes into more concise and organized summaries, a process we call *incremental diagenesis*. The provided authoring tool permits easy creation and modification of the summaries, and allows *incremental summarization*, a process in which summaries are incrementally constructed and distilled. Arkose is augmented with information retrieval mechanisms and visual aids to help editors quickly identify important topics and relations among posts, authors of those posts, and summaries. We believe that Arkose will be useful at distilling other types of informal information as well.

The paper proceeds as follows. First, we discuss how this work fits into related HCI and CSCW research. Then, we discuss the problems of a typical discussion forum and the need for distillation in more detail. This is followed by a distillation scenario using our Arkose system to present some of the main features provided. We then discuss the design principles upon which Arkose is built and end with the technical details of the system.

GROUP'07, November 4–7, 2007, Sanibel Island, Florida, USA.

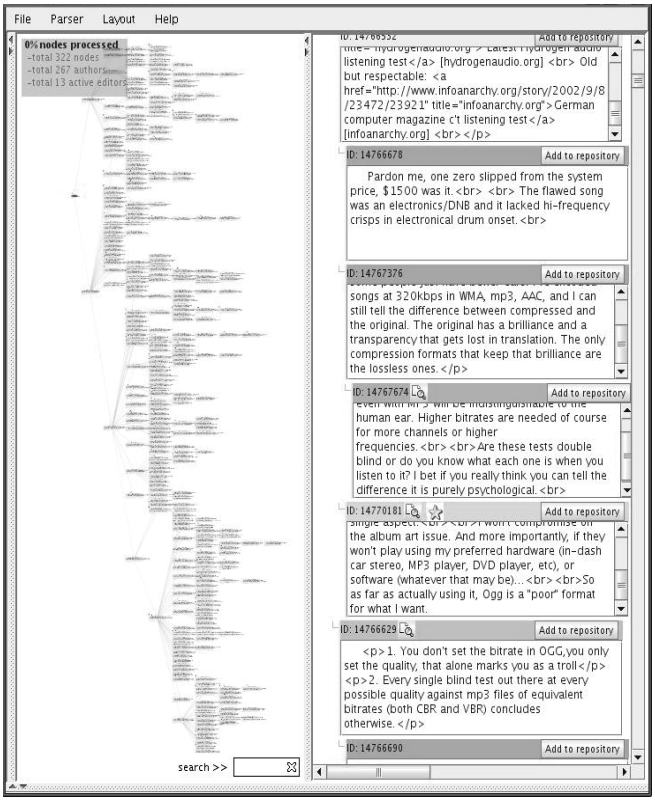

Figure 1: An overview of the navigator. The left column graphically presents the discussion space in a tree structure, which also functions as a substrate for distillation. The tree is fully zoomable and draggable with an online search capability for easy navigation. The right column displays each post in a more web forum like interface for maximum readability.

2. RELATED WORK

This project builds on three streams of HCI and CSCW research. The first is the design rationale, an area of considerable interest in the late 1980s and early 1990s. The hope was to reuse design and decision understanding. The design rationale research stream (e.g, [Moran and Carroll 1996], [Buckingham Shum 1996]) examined both languages and representations (e.g., [Lee 1990], [MacLean et al. 1990]) as well as interfaces (e.g., [Conklin 1992]) for supporting design history and explication. Concomitant with this interest were field studies of actual use. The systems failed to gain widespread usage. The leading cause for this lack of use was found to be that their use required conscious and slow activity on the part of a group creating the design, especially for formalizing the activity while it was occurring [Buckingham Shum 1996]. The support systems for this did not remove this effort, and it interrupted the natural flow of activity. In addition, as Grudin [1996] pointed out, it required considerable work to create a design rationale, and this was work that had an unclear payoff for those creating it. In other words, design rationale systems interrupted normal social behavior, and they had an unclear set of incentives for use.

Similarly, argumentation support systems were studied in AI (e.g., [Hurwitz and Mallery 1995]) and decision support and rationale systems were studied in Information Systems, with similar results.

The second line of HCI research is the considerable work in using online communities to garner knowledge. Knowledge communities and communities of practice have been well studied [Preece 2000] [Wenger 1998]. Everyday, people put significant effort into online discussions (e.g., with Slashdot as in [Lampe and Resnick 2004]). The WorldJam experiment put IBM employees from around the world for a limited time into an online community for brainstorming [Millen and Fontaine 2003]. Our earlier work includes iDIAG/Forum, a prototype for creating community brainstorming [Ackerman et al. 2003].

Recently, there has been interest in understanding how to obtain knowledge artifacts from online community interactions. The best studied form is the wiki. Wikis are a form of online community, albeit a specialized form. Wikipedia has been enormously successful, but there have also been many failures. Only some results from online community discussions are suited to wikis (compendium entries were a primary form found by [Hansen 2007]). Our usage scenarios, with the distillation of online discussions, are not necessarily suited for wikis. Hansen et al. [Hansen et al. 2007] examines some of the issues in tying an online community with a wiki.

As such, a common problem in online discussions, the issue of ending up with an information space that is hard to maintain and reuse, must be addressed for online discussions to be of use. Topics and their discussions tend to be scattered around incoherently, and search and comprehension of important and relevant information are sometimes problematic. One approach to the problem has been explicit structuring and organizing of online collaborations as they occur. As mentioned, earlier design rationale and argumentation support systems focused on providing formal node and link grammars for participants to structure the discussion discourse (e.g., gIBIS [Conklin and Begeman 1988] and Aquanet [Marshall et al. 1991]). Recent approaches in online policy discussions include Farnham et al. [2000] which also explicitly structured online discussions to improve computer mediated decision making.

These approaches provided well-defined structures for a discussion discourse, but the imposed formal structure limits how participants can discuss topics. Most online forums do not place restrictions on how conversations should progress. A thread on an online discussion forum is initiated as needed, and it often forks into other related (or sometimes unrelated) topics. The free nature of online discussions allows a naturalness of interaction, and is likely to be one of the reasons online forums have flourished. Thus, imposing formal rules and structure in discussion raises barriers for participants. Famham et al. supports this point in their findings that an explicit structure imposed in conversation is interruptive and restricting.

The issue of imposing formal structure is exacerbated when the participants are required to make an upfront decision about what the information structure should be. Incremental formalization [Shipman and McCall 1994; Shipman and Marshall 1999] as an approach allows information to be gradually formalized over time. (The authors define formalization as “the process of identifying machine-processable aspects of information”, but in general it includes allowing intermediate levels of structure and use of formal representations.) This reduces users’ pressure to initially commit to a specific format and organization. Building on this idea, our approach in distillation of a large informal information

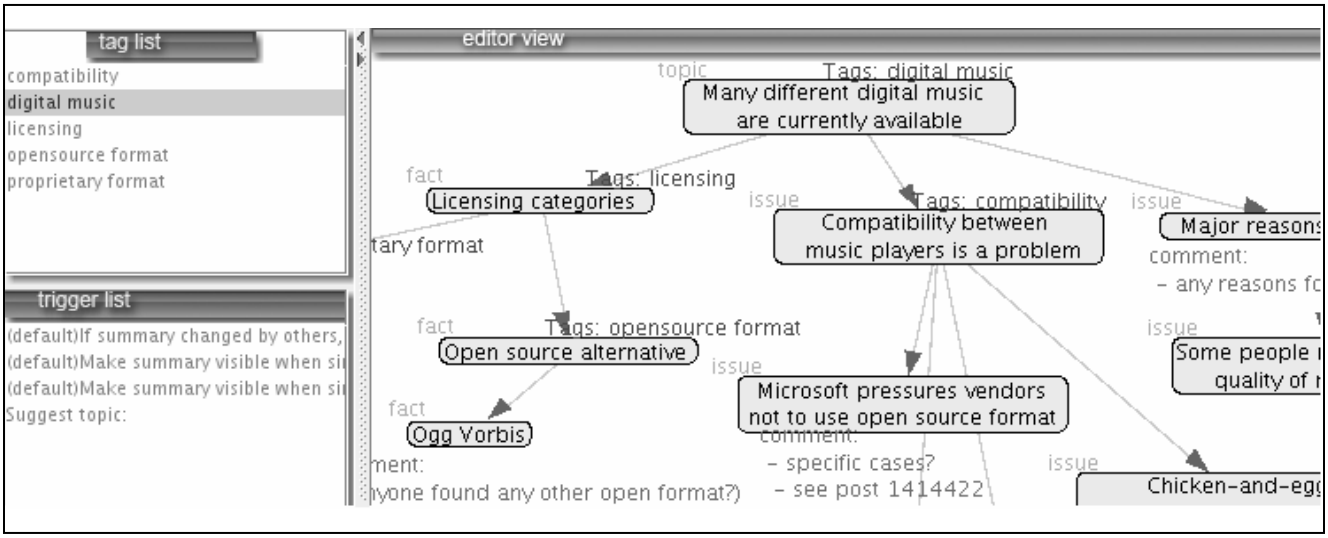

Figure 2: A partial overview of the authoring tool. An editor can create a summary and modify an existing one. Each node is associated with an editor customizable type. Notice that there are "meta-discussions" taking place (noted as "comment"). The trigger list (bottom left column) displays user specified trigger conditions, which are used to notify the editor when an associated event occurs. The graph is fully zoomable and draggable, with creation and deletion capabilities of nodes and links. An editor can also import multiple existing summaries and merge them.

space allows a gradual increment in the organization and reusability of an online community's information.

2.1 The need for distillation

While the aforementioned free nature of most online discussion forums may have contributed to their popularity, at the same time, it has been one of the reasons a typical discussion space is left unorganized and unstructured. Consequently, there is a greater need for an explicit distillation process during and/or after the discussion. By providing distillation facilities separate from the discussion itself, participants in the online discussion can freely perform discussions without worrying about structuring the discourse.

To do this, several problems must be addressed and support for their solution or amelioration is required. They include:

1) As previously mentioned, discussion spaces in general are hard to comprehend. This is especially true after the discussion space has grown considerably with dozens of topics and hundreds or thousands of posts. Interesting topics and ideas are often scattered around and buried deep in discussions, and are not easy to locate unless the user reads a considerable amount of posts.

2) Some topics in the discussion space are duplicates. This may or may not be deliberate; users may not realize a topic has been discussed in another part of the discussion space and start a new thread. Users may intentionally start a duplicate thread to push their ideas. In either case, duplicate discussion threads waste users' time and effort spent in the discussion space, and they only compound the incomprehensiveness of the space discussed in 1).

3) Some topics will be socially problematic, controversial, or off-topic. These posts make it harder for users to participate effectively.

Even if users adhere to a strict code of conduct, and all the discussion topics and posts are meaningful and valid, the discussion space will still need some organizing for reuse. For example, a policy discussion at the end needs to report people's points of view for policy makers. The current shape and form of discussion forums are not suitable for such a report.

Our approach to the problems is through *distillation*. By distillation we mean the process of creating a more concise and organized form of information. It is more than just text summarization; rather, it includes sorting through large corpus of text for interesting topics, pruning away redundant and off-topic discussions, identifying interesting authors, different points of view, and ultimately making the information more reusable for later purposes. Ideally, the distillation process would be performed by a variety of natural language processing methods to minimize the use of precious human resources. Techniques such as automatic summarization [Radev and Hovy 1999], discourse analysis [Passonneau and Litman 1997], and sentence structure parsing [Klein and Manning 2001] have been gradually improving, but they do not yet provide the human level cognitive abilities necessary for the aforementioned distillation tasks. Thus, we view distillation as a system augmented process, guided and directed by experienced human editors, rather than an entirely automated one.

We next turn to a usage scenario that will be used to illustrate what we believe are the necessary design principles for distillation support.

3. DISTILLATION SCENARIO: A DISCUSSION ON THE FUTURE OF THE UNIVERSITY

3.1.1 Setting

The basic scenario for Arkose's use is that a university has recently held an online discussion forum on various topics on the future of the university. The forum was open to the various

interest groups in the university to reflect different views and ideas. After two weeks of lively discussions on numerous topics, the university temporarily closes the forum and commissions four employees to distill the discussion space. Jack is assigned as one of the four editors. While residing in the same building, the editors are physically separated into their respective offices. In addition, not all of the editors can work on the report at the same time due to schedule conflicts. Any number of editors from one to four would simultaneously use Arkose to perform distillation tasks.

3.1.2 Visually supported

Jack runs Arkose and finds the online discussion space presented in a graphical tree form in the navigator (Figure 1). He sees by the instant message that other editors, Ken and Lisa, are currently online, while Matt is offline. Initially, the discussion space is completely zoomed out as an overview. Using a mouse, Jack zooms in and drags the tree around to scan through the posts. He quickly looks at a few top nodes of the tree, as it is often the case that the top nodes contain main topics.

Jack can also get a sense of the topics in the space without reading the entire content of the posts; Arkose has automatically found keywords from the discussion space and attached them to the relevant posts. Jack is distilling a specific part of the discussion space, and that is visually indicated with colored aggregates of the posts with tags (Figure 4). After reading a few posts in a thread, Jack finds an interesting discussion taking place. He decides to distill this part of the discussion space, and selects the posts and tags them as "Being worked on". This creates a yellow aggregate around the posts and reduces the size of the posts to inform others of the status (Figure 4, part 1).

3.1.3 Initial Summary Creation

The creation of an aggregate automatically copies the posts into the authoring tool (Figure 2) with which Jack creates a summary structure. He creates a topic node and a few subsequent "issue" nodes and "fact" nodes, and links them together. He also assigns tags to some of the nodes in the summary structure that reflect their contents. But before being able to finish the entire summary, Jack realizes he has a meeting in five minutes and stops the process. He exports his summary and tags it as "Initial work started" and "Do not modify". This changes the color of the aggregate to a light green (Figure 4, part 2) and informs the editors of the summary's status.

3.1.4 Incremental summarization

After coming back from the meeting, Jack resumes the process. He imports the previous summary into Arkose's authoring tool, and continues working on the summary structure. At this point, he feels some other editor with more expertise in the domain area should review the summary he has created. He leaves a question on the node, and exports the summary as "Attention needed" (Figure 4, part 3). This changes the aggregate color to red and enlarges the posts to give more visibility. The requested information is appended into a message area in Arkose where everyone can read it. Another editor, Lisa, sees the request and decides to help Jack by adding some reference points to the question. Jack is then notified of this, and comes back to this summary to complete it. After Jack feels the distillation is finished, he exports it as "Closed" (Figure 4, part 4). This is indicated as a blue aggregate with the original posts reduced in size to indicate there is little need to reopen them (however, an editor can reopen any part of the discussion space if necessary). This replacement of the original posts with a summary node not only condenses the discussion space, but it also makes the information more organized and understandable. Jack then moves on to another part of the discussion space to continue distillation.

3.1.5 Author Network

Jack has found an author with a very interesting point of view. First, he moderates the post up with a star symbol (Figure 4) to indicate the node has insightful information. He is interested in finding out more about what the author has discussed, and opens up a visual aid, "Author Network" (Figure 6). The Author Network visualizes the conversational activities among the authors. Jack can look at the keywords and contents of the posts between the author and others to quickly grasp the topics.

3.1.6 Merging summaries

Meanwhile, Ken has been working on his summary without realizing there already was a summary with similar content. Obviously, there were two discussion threads on the same topic, and the editors did not realize this because the threads were physically apart. The information retrieval mechanisms within Arkose not only tries to find duplicate posts within the discussion space, but it also compares a newly created summary with existing ones (details below). As Ken is creating the summary, Arkose notifies him of the possible existence of a similar summary. Ken reopens the suggested summary and its original discussion thread, and finds that it would be better if the two were combined. Ken then imports both of the summaries and merges them together. Some of the parts are duplicates and are deleted. Some other

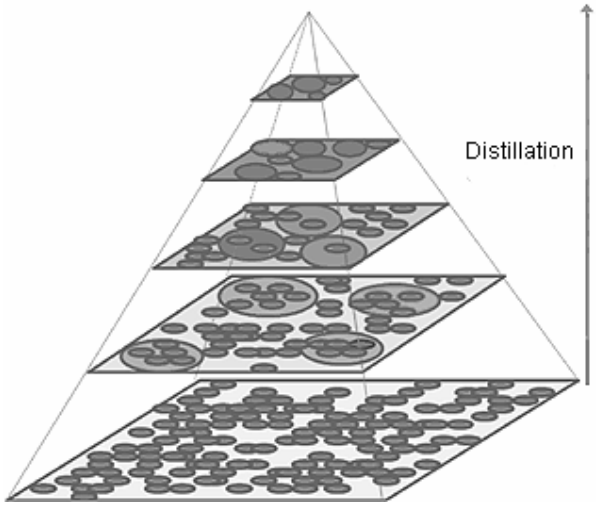

Figure 3: The conceptual view of Incremental Diagenesis. As distillation progresses the unorganized bramble of discussion space (depicted as the bottom plane) is gradually replaced by summaries and other meta-information. The small circles represent discussion posts while the aggregate represent some stage of distillation process.

nodes are linked together to form a bigger and more complete

summary. Ken exports the new summary out to the navigator. Since Jack's summary has been changed by another editor, Jack is notified of this.

3.1.7 Keyword farm

Matt, who was initially offline, now joins the other editors. Rather than starting a new summary, he decides to work on existing ones and imports one of the summaries created by another editor but not completed yet. After some more distillation work he exports the summary with a "Closed" tag to indicate the summary is complete. The status panel of Arkose shows that 10 % of the discussion space has now been distilled. At this time, Arkose compares the keywords within the created summaries with the keywords of the discussion space it found in the beginning of the user's session. Some of the keywords that Arkose thinks are important have been used in the summaries and tags; while others have not. Arkose notifies the editors about these unused keywords in "Keyword Farm" (Figure 5). Keyword Farm visually presents keywords with two types of information: one is the machine calculated probable importance score of each keyword, and the other is the actual usage data of each keyword in editor-created summaries and tags. Matt can easily tell the type of distillation process in which each keyword has been used. He reads the posts associated with suggested keywords in Keyword Farm, and decides a summary indeed needs to be created around some of the keywords. He then looks at the discussion threads on the keywords and proceeds with the distillation process.

3.1.8 Ending distillation

After one week, the distillation process has reached an end. The original discussion space in the navigator has been transformed into mostly completed summaries (represented as blue aggregates in Figure 4, part 4), some partially processed summaries (represented as green aggregates in Figure 4, part 2), and other meta-information such as editor-created tags, comments, and question-and-answers. The summaries are stored in an XML format. The collection of the editor-created summaries is presented with a style sheet format such as CSS or XSL, to form a report of the discussion space.

3.1.9 Reuse

The summary report is forwarded to the university's policy makers. It contains the important topics and their arguments in an orderly fashion. It also shows a list of topics that have been discussed sufficiently, as well as topics that did not generate enough or sufficient discussion. These have been identified by the editors through the distillation process. The policy makers decide to reopen the discussion forum for another few days to mainly discuss the insufficiently discussed topics. This time, the discussion space not only contains the original threads, but also attached to them are the editor-created summaries and meta-information. Returning discussion participants do not have to read the entire posts again to understand the topic; rather, they could read the attached summaries and meta-information. Authors are invited to join the open topics according to their associated keywords computed in Author Network. After a few days of discussion, the forum is again closed and distillation of the newly added information once again takes place.

4. TWO DESIGN PRINCIPLES

In the previous section, we looked at a distillation scenario in detail. Many of the required features are enabled by two design principles upon which Arkose is built. These principles set Arkose apart from previous systems.

4.1 Incremental summarization – Allowing intermittent states

One feature lacking in earlier design rationale systems is allowing users to not specify up front what the information structure should look like. In other words, later changes to the existing structure or schema were either not supported or difficult to do in earlier systems. Our approach to the problem follows from Shipman and McCall's incremental formalization. As mentioned, incremental formalization allows information piece to be gradually formalized. Thus, it adds much flexibility in creating and modifying information structure and takes a burden of initial commitment off the users. While our main goal in distillation is not the formalization of the information (into, say, semantic web statements), we adopt the idea to support a flexible distillation process, which we call *incremental summarization*. Its advantages are:

- *Low overhead cost.* The idea of incremental summarization is simple; yet, it has many ramifications in the way editors perform their tasks. An editor has the capability to modify, extend, and merge or divide existing summaries. The editor does not have to worry about the final structure of the summary, thus much less coordination with other editors is needed especially about the format or structure of the corpus. This reduces the overhead cost of initial structuring of a summary.
- *More thorough summaries*. Topics related to each other may be discussed separately in different places in the discussion space. Since they are separated, an editor may not realize there are more discussion threads on the topic being summarized and later discover them after having finished a rudimentary summary. As part of incremental summarization, an editor can extend an existing summary with newly found topics and evidence, making the summary more complete.
- *Better expertise distribution*. Another way incremental summarization may help is by better distributing editors' expertise. Each editor may have a different level of expertise in any given subject area, and that may affect the quality of the summaries the editor creates. An editor can tag a summary as incomplete and attach a comment that asks for help in a specific area. This information would then be listed where it is visible to everyone. The editor may, of course, directly ask a particular editor through an instant message or email if the required expertise is known. Thus, the summary can be incrementally summarized, allowing a more effective expertise distribution.

4.2 Incremental diagenesis – Bringing order to chaos

The second design principle comes from the fact the distillation process starts from a large unorganized information space and ends with a smaller and tighter summary space; a process we call

incremental diagenesis. Diagenesis is the conversion of sediment into rock, connoting a loosely scattered large amount of information being transformed into a more concise and organized state. This is quite different from earlier systems in that most of their processes start from an empty space. In the previous distillation scenario, the discussion space has been used as a substrate for distillation. This transformation of the space is a gradual process where at any given time the space consists of heterogeneous information entities: the original posts, meta-information such as editor assigned tags and comments, post scores and keywords, and various stages of editor-created summaries of the discussion threads.

Figure 3 shows the conceptual view of the transformation of the information space through distillation. The small circles represent individual posts or discussion threads that have not been distilled. The bigger encompassing circles indicate that some distillation process has been applied. Different colors imply various stages of distillation. The crosscuts of the pyramid represent the stages of the information space as distillation progresses from bottom to top. At the beginning of distillation (depicted as the bottom plane), the space is essentially raw discussion data imported from an online discussion forum. At this stage, the space is less organized and contains some duplicates and information with little value. It may consist of dozens of discussion threads with hundreds to thousands of posts attached to them and meta-information such as moderation scores or comments. The existing meta-information is presented as notes attached to appropriate places in the space.

As distillation progresses, more and more raw discussion data are transformed into distilled summaries. (Note that different topics can be at different levels. One topic may be completely distilled while another is still intermediate.) This reduces the size of the information space (as depicted at the top of the pyramid), with the information more organized and structured. In the previous distillation scenario, the discussion posts have been replaced with a smaller summary structure. The initially unorganized space gains order and reusability through the distillation process. The process is fully visible in the navigator (a zoomed in partial view of the progress is shown in Figure 4).

Both incremental diagenesis and incremental summarization help editors flexibly distill informal information. Arkose is built as a demonstration of these principles, providing tools and mechanisms that allow a gradual increase in the organization of the informal information.

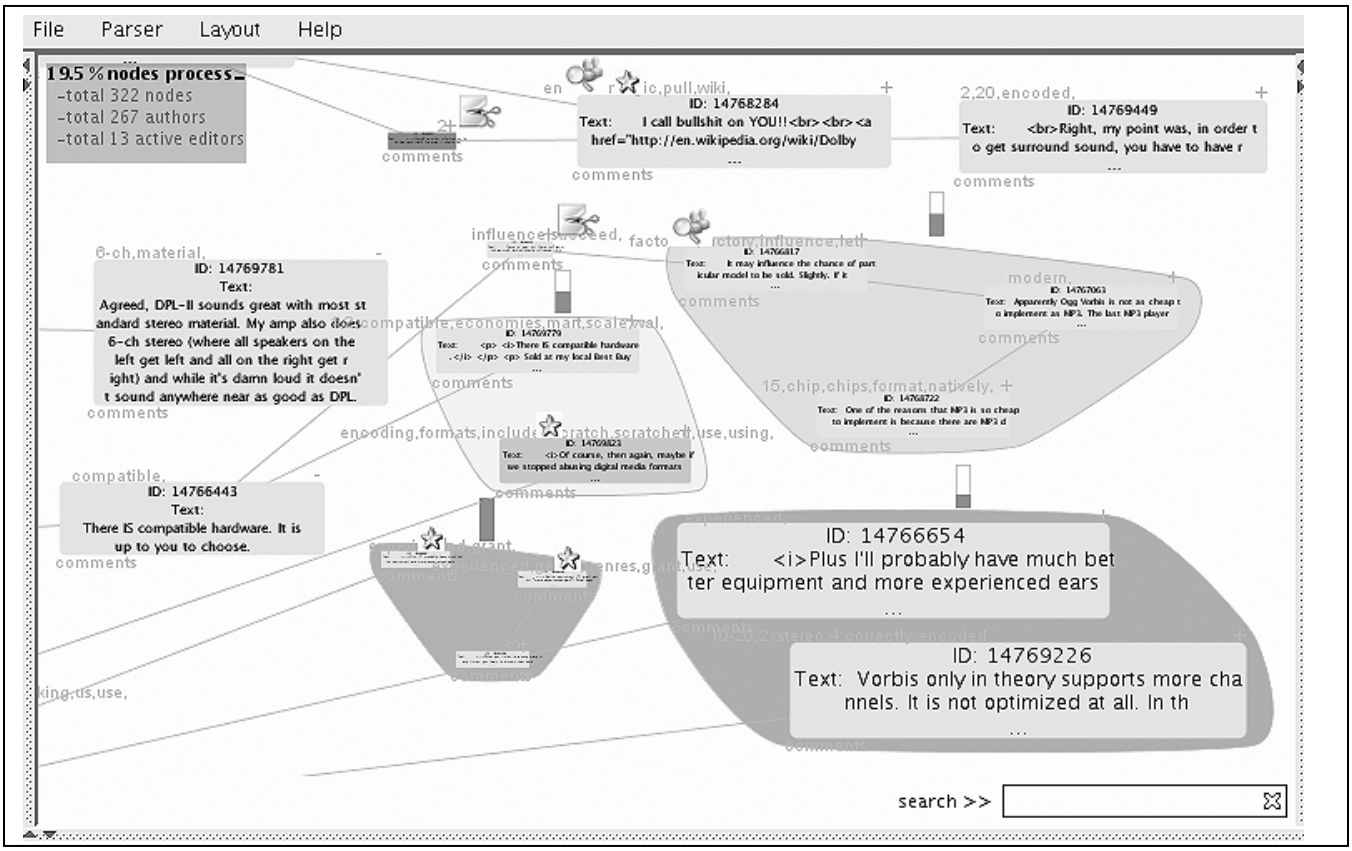

Figure 4: Incremental diagenesis in progress. This is a zoomed in view of the navigator (Figure 1). Each aggregate represents a different stage of distillation. Posts and summaries have varying visibility (represented in their sizes and colors) according to their status. Note that the red "Attention Needed" aggregate has bigger sized post while the blue "Closed" (or completed) aggregate has reduced sized posts. Tags and comments can be directly left on the space. Gradually, the original discussion space is transformed into distilled summary outcomes. (The text "Part 1" through "Part 4" are for reference purposes only.)

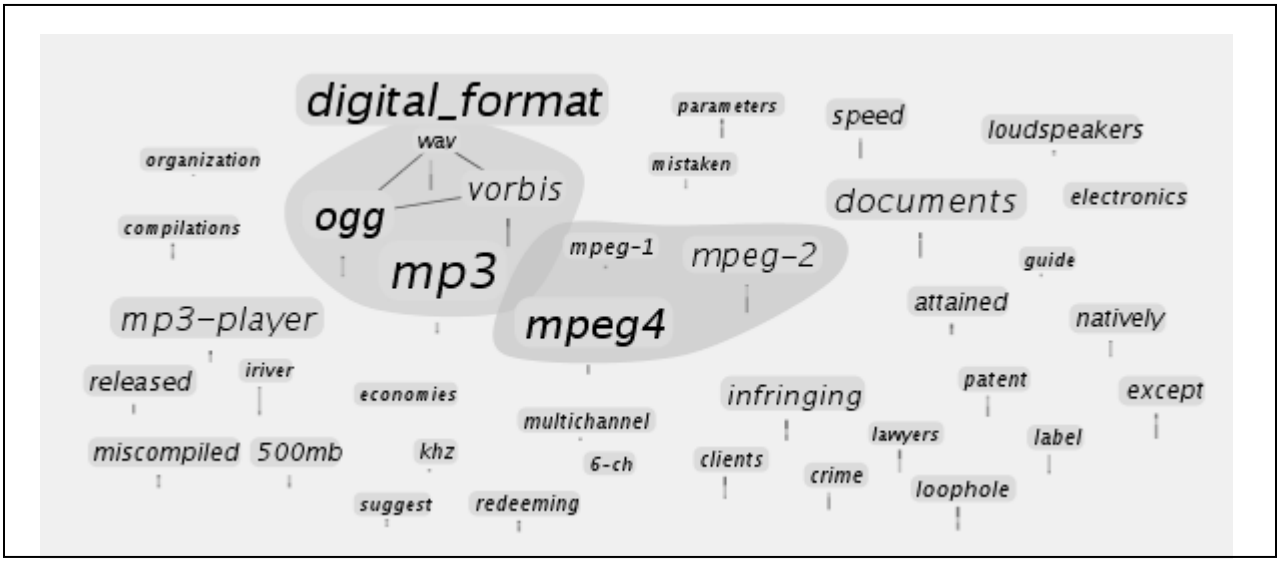

Figure 5: A partial view of the Keyword Farm, a visualization of keywords in the discussion space. This visual aid helps editors quickly understand the topic space and allows them to rearrange and group the keywords. It can suggest editors of possible important topics that have not been covered yet.

In the previous sections of the paper, the details of the navigator, authoring tool, and visual aids that constitute Arkose have been purposefully left out to concentrate the presentation on the distillation process and its requirements. We now discuss the technical details of Arkose in the next section.

5. TECHNICAL DETAILS OF ARKOSE

Arkose consists of approximately 12,000 lines of Java code, along with with the Swing user interface toolkit and the Prefuse toolkit [Heer 2007]. Prefuse provides a rich set of visualization and interaction features with animation, search, and database connectivity.

Arkose itself is developed to be discussion forum agnostic. In other words, any text based online discussion forum can by used by Arkose, provided a parser converts the forum into the XML-based TreeML format [Fekete and Plaisant 2003]. One of the forums supported is the iDIAG/CyberForum system described in [Ackerman et al. 2003].

Arkose consists of four major components: the navigator, the authoring tool, and visual aids. Each is covered in turn below.

5.1 The navigator

The navigator is relatively straightforward, but necessary. The purpose of the navigator (Figure 1) is two-fold. One is to visually present the original discussion space in order to solve the partial-view problem inherent in existing web based forums. This helps editors better understand the discussion space by providing an overview and allowing them to focus on specific parts as necessary. The discussion space also serves as a substrate for the distillation process. Distilled summaries, tags, various forms of meta-information, and comments are added to the space as the distillation progresses, gradually transforming the space into a more organized and reusable state. The navigator makes all the activities of editors visible for a more effective collaboration. The discussion space is fully zoomable and draggable and provides online search capability of the content. The navigator supports multiple visibility levels of posts according to their current distillation status.

5.2 The authoring tool

As with the discussion space in the navigator, individual summaries are also incrementally shaped. An editor creates and modifies summaries of the discussion space in the authoring tool (Figure 2). As mentioned, a summary may be updated as many times as needed by different editors until it is deemed to be complete. Adding and deleting to and from a summary is as simple as connecting and disconnecting edges in the summary graph. This also allows easy merging of multiple summaries.

Each node in a summary graph is a typed entity that indicates its role. For example, in the example in Figure 2, the types are “topic”, “issue” and “fact”. These types are customizable, allowing editors to add new ones and modify existing ones. When an editor needs to create a new type, one can be created without any restrictions. However, when an existing type is modified (for example, changing “fact” to “evidence” to better reflect the role), the editor is asked to specify an explanation or justification for the change, which can be read by other editors later.

Each summary is assigned a distillation status when it is exported back into the discussion space in the navigator. The original posts are aggregated with a summary. Currently, there are four types of status indicating the distillation progress: “Being worked on”, “In Progress”, “Attention needed”, and “Closed.” A summary with the “Being worked on” status indicates that it has been imported into one of the editors’ authoring tool. This tells other editors that they should not modify it. “In Progress” means that initial work has been done, and the current work is exported back to the discussion space. An editor may freely import the summary and work further on it, in which case the status becomes “Being worked on” to prevent other from working on it. A summary can also have a more detailed progress indication. Since a summary

may be revisited many times before its completion, some indication of its relative progress might be helpful. An editor can specify this by dragging a slider to set how full a bar icon is. At each revisit, the editor would raise up the slider to fill the bar. When the summary is completed, the bar would be fully filled. The idea behind having this kind of secondary indication is to help editors quickly understand the status of summaries in progress. This way, the number of colors for different summary statuses can be kept minimal (currently four) while providing richer information. "Attention needed" is usually left with an editor comment from the editor. This may be a question or concern that requires another editor's help. "Closed" indicates the summary is complete, but Arkose allows editors to work on it further if they wish.

The bottom left column of Figure 2 shows the "trigger condition" entries. A trigger condition is a rule used by an automatic process in Arkose. There are currently three default trigger conditions implemented; 1) notify the editor when another editor modifies the summary, 2) notify the editor when a summary with similar tags is found, and 3) notify the editor when a summary with similar node content is found. The content similarity is calculated by the cosine similarity of term vectors of the nodes in the summary.

As the editor works on a summary, the automatic process compares its content and tags with existing summaries to find whether similar summaries already exist so as to minimize duplicate work. Currently, only three pre-defined triggers are available. We are currently investigating ways to allow editors to write simple menu-driven scripts to specify various trigger conditions. Allowing editors to post and share these customized scripts would facilitate stronger coordination.

5.3 The use of visual aids

As well, Arkose consists of several visual aids that are critical to supporting incremental summarization and incremental diagenesis. These include a visualization of keywords in the discussion space that we call the "Keyword Farm", and a visualization of conversational activities of the post authors that we call the "Author Network". The following subsections discuss these visual aids in detail.

5.3.1 Keyword Farm

It is often helpful to quickly see the keywords used in an information space to grasp what topics are being discussed. Recent social bookmarking sites such as del.icio.us implement new ways of utilizing meta-information by displaying user created tags in such a way that visibility varies according to the importance of the tags. We have created a similarly purposed visualization, the "Keyword Farm" (figure 5). The Keyword Farm visualizes selected keywords from the discussion space so

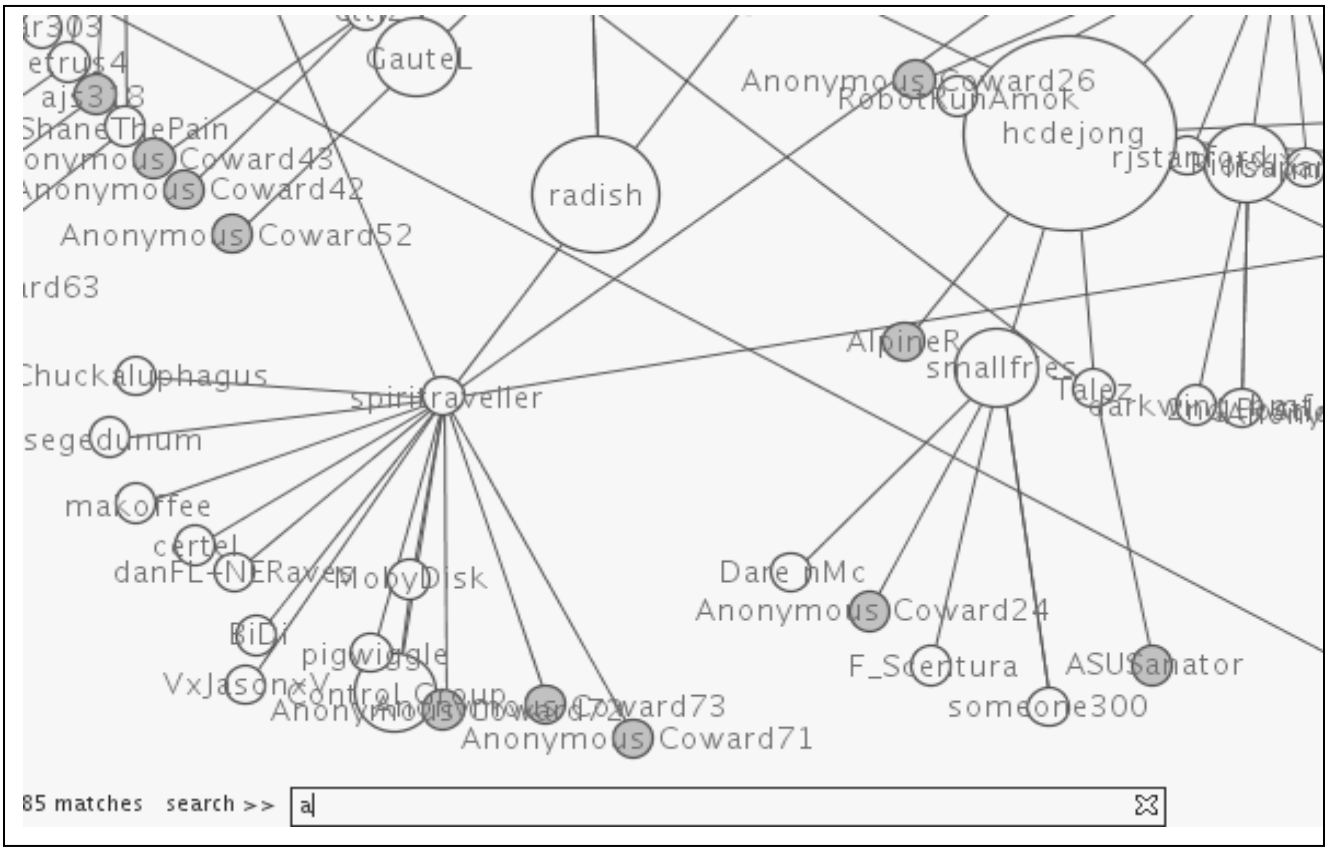

Figure 6: A partial overview of the Author Network. The size of the circle indicates the number of posts a particular author of the discussion space has left. The edges indicate who replied to whose post, with keywords with high tf-idf values that are common to both authors (when clicked). Keywords found in the posts are associated with respective authors.

editors can more easily see what needs to be done.

First, we have integrated WordNet [Fellbaum 1998] using the MIT Java WordNet Interface to help identify synonyms to allow editors to easily build a quasi-ontology. An editor may replace a number of semantically similar keywords with an overarching new word, or specify a relationship between two keywords.

Second, the visualization utilizes both machine calculated data and actual usage data of the keywords to give editors some useful functionality. One is a standard information retrieval technique to calculate a word's probable importance value, or the term frequency times inverse document frequency (tf-idf) to give editors statistically computed keywords. The top *p* keywords from the original discussion space are picked according to the following:

> First, the words in each post are tokenized, filtered with a stop word list, porter stemmed [Porter 1980], and form a term vector for the post. After all the unique words are gathered from the entire nodes, each word's term frequency (how many times the word appears in the entire space) times inverse document frequency (inverse of the number of documents in which the word appears) is calculated.

The tf-idf scores are visually represented as the size of the words in the Keyword Farm. Thus, the bigger the word is the more likely it is significant. In addition to showing the importance values of individual keywords, a matrix of word frequency is used to indicate groups of words frequently appear together in the discussion space. Upon selecting a keyword in the Keyword Farm, lines visually connect the word and its associated words to form a graph with the frequency information presented as the varying thickness of the lines. In addition to the machine calculated keywords, editors can add a new keyword or delete an existing one.

The Keyword Farm also suggests important topics that have not yet been covered by the editors, providing for incremental diagenesis. This is done through the second type of information represented in the Keyword Farm, which is each word's actual usage in summaries and tags. The graph bar under each word indicates the word's actual usage. As editors organize the discussion space and create summaries of discussion threads, some of the keywords (raw or as described in the quasi-ontology) are included in them. By visually presenting the usage information, editors can have a better understanding of the work progress, and identify topics that are sufficiently covered and topics that need further organization. In the Keyword Farm, all the keywords start from the ground initially and as they are used by editors the graph bars under the keywords grow taller so as to push the keywords upward. Colored sections in the bar indicate specific usage cases. For example, the top blue portion represents how frequently the word has been used in a topic in a summary, the second green portion represents the frequency with which the word has been used in editor-specified tags in various places in summaries.

After distillation has progressed to some degree (the current threshold is when every 10% of the posts are distilled), a number of keywords have been used and thus their graph bars in the Keyword Farm have grown accordingly. However, the Keyword Farm may notice that some of the keywords that it thinks are important (i.e., words that have high tf-idf scores) have been used very little in summaries and tags. This may indicate that editors simply overlooked these topics, the portion of the discussion space has not been distilled yet, or the tf-idf values for the keywords do not correctly represent their actual importance. When notified by the Keyword Farm, an editor may check to see whether the suggestion is a valid one. The editor could then initiate a new distillation process over the discussion space where the keywords are relevant, or the editor may simply turn off the suggestion if it is not correct. Thus, not only more value is gradually added to the Keyword Farm, but it also helps editors incrementally organize the discussion space.

5.3.2 Author Network

Another supportive technique is visually presenting the conversation activities of the authors in the discussion space. One of the goals of distillation, as discussed earlier, is to identify authors with interesting ideas. If is often helpful to know how active a particular author is and what the author's messages are about. The Author Network visualizes the information in a social network, where an editor can search for authors and their discussion contents and keywords. Each circle in the network represents an individual author. The size of the circle indicates how many posts the author has written, thus showing magnitude of the author's activity, and the links or edges between authors in the visualization represent conversation activities. Keywords from their conversations are visually attached on an edge, so an editor can quickly scan through what the authors talked about. As the distillation progresses and editors identify authors with interesting ideas, tags can be added to emphasize those authors.

6. CONCLUSION

We plan to evaluate Arkose among a small group in near future to examine the usefulness of the Arkose system and to improve it further. We will also explore possibility of extending Arkose to distill other domains. This paper has presented the need and requirements for *distillation* to reuse informal information, especially from an online community's brainstorming and discussions. We have presented two design principles *incremental summarization* and *incremental diagenesis* that allow a more flexible distillation process. On these design principles we developed Arkose, a system with a set of augmentative tools for supporting incremental diagenesis and incremental summarization, to support human editors in collaboratively handling this informal discussion information.

7. ACKNOWLEDGEMENTS

This work has been funded, in part, by the National Science Foundation (IRI-9702904) and the University of Michigan/School of Information Alliance for Community Technology. We would like to thank Dan Atkins for the original scenario of use, as well as Cliff Lampe, George Furnas, Paul Resnick, Jun Zhang, Xiaomu Zhou, Jina Huh, and Ben Congleton for their insights and help with this work.

Arkose continues the iDIAG project. [Ackerman et al. 2003] discusses an earlier version of this project; this is the second generation of the distillation system.

8. REFERENCES

[1] Ackerman, M. S., and McDonald, D. W. 1996 Answer Garden 2: Merging Organizational Memory with Collective Help,

ACM Conference on Computer-Supported Cooperative Work (CSCW'96), pp. 97-105.

[2] Ackerman, M. S., Swenson, A., Cotterill, S., and DeMaagd, K. 2003. I-DIAG: From Community Discussion to Knowledge Distillation, *International Conference on Communities and Technologies*.

[3] Buckingham Shum, S. 1996. Design Argumenttion as Design Rationale, *The Encyclopedia of Computer Science and Technology*, pp. 95-128. New York: Marcel Dekker, Inc.

[4] Conklin, J. 1992. Corporate Memory, *Groupware '92*, 131-137. San Jose: Morgan-Kaufmann.

[5] Conklin, J. and Begeman, M. L. 1988. gIBIS: a hypertext tool for exploratory policy discussion. *Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work*, 140-152.

[6] Fekete J. and Plaisant C., "DTD describing a tree structure for visualization," 2003, http://www.nomencurator.org/infoVis2003/ download/treeml.dtd> (26 May 2007).

[7] Fellbaum C. (editor). 1998. *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

[8] Grudin, J. 1996. Evaluating Opportunities for Design Capture. In T. P. Moran & J. M. Carroll (Eds.), *Design Rationale: Concepts, Techniques, and Use*. Mahwah, NJ: Lawrence Erlbaum Associates.

[9] Hansen, D. L. 2007 Knowledge Sharing, Maintenance, and Use in Online Support Communities: Ph.D. thesis, University of Michigan.

[10] Hansen, D. L., Ackerman, M. S., Resnick, P. J., and Munson, S. 2007. Virtual Community Maintenance with a Collaborative Repository, *Annual Meeting of the American Society for Information Science and Technology (ASIS&T)*, in press

[11] Heer J. 2007. *"prefuse | interactive information visualization toolkit,"* 05 April 2007, <http://prefuse.org> (26 May 2007).

[12] Hurwitz, R., and Mallery, J. C. 1995. The Open Meeting: A Web-Based System for Conferencing and Collaboration, *4th International WWW Conference*.

[13] Klein, D. and Manning, C. D. 2001. Parsing with treebank grammars: empirical bounds, theoretical models, and the structure of the Penn Treebank. *Proceedings of the 39th Annual Meeting on Association For Computational Linguistics*, 338-345.

[14] Lampe, C. and Resnick, P. 2004. Slash(dot) and burn: distributed moderation in a large online conversation space, *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems*, 543-550

[15] Lee, J. 1990 SIBYL: A Tool for Managing Group Decision Rationale, *Proceedings of ACM CSCW'90 Conference on Computer-Supported Cooperative Work*, 79-92

[16] MacLean, A., Young, R., Bellotti, V., and Moran, T. 1990. Questions, Options, and Criteria: Elements of a Design Rationale for User Interfaces: EuroPARC/AMODEUS.

[17] Marshall, C. C., Halasz, F. G., Rogers, R. A., and Janssen, W. C. 1991. Aquanet: a hypertext tool to hold your knowledge in place. *Proceedings of the ACM Conference on Hypertext*, 261-275.

[18] Millen, D. R., & Fontaine, M. A. 2003 Multi-team facilitation of very large-scale distributed meetings *European Conference on Computer Supported Cooperative Work (E-CSCW 2003)*, 259-275

[20] Moran, T. P., & Carroll, J. M. 1996 Design Rationale: Concepts, Techniques, and Use. Mahway, NJ: Lawrence Erlbaum Associates.

[21] Passonneau, R. J. and Litman, D. J. 1997. Discourse segmentation by human and automated means. *Comput. Linguist.* 23, 1 (Mar. 1997), 103-139.

[22] Porter M. F., 1980. An algorithm for suffix stripping. *Program*, 14(3), 130-137

[23] Preece, J. 2000 *Online Communities*. New York: Wiley.

[24] Radev, D. R., and Hovy, E. 1999 Intelligent Text Summarization - Report on the AAAI Spring Symposium. AI Magazine, 20(3).

[25] Shipman, F. M.. and Marshall, C. C. 1999 Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Computer Supported Cooperative Work Journal*, 8(4), 333-352.

[26] Shipman, F. M. and McCall, R. 1994. Supporting knowledge-base evolution with incremental formalization. *Proceedings of the SIGCHI Conference on Human Factors in Computing System*. 285-291.

[27] Wenger, E. 1998. *Communities of practice : learning, meaning, and identity*. New York: Cambridge University Press.