ScienceSifter: Facilitating Activity Awareness in Collaborative Research Groups through Focused Information Feeds

Linn Marks Collins[1], Ketan K. Mane[1,2], Mark L.B. Martinez[1], Jeremy A.T. Hussell[1], and Richard E. Luce[1]

[1]*Research Library, Los Alamos National Laboratory, Los Alamos, NM 87545 USA*
[2]*School of Library and Information Science, Indiana University, Bloomington, IN 47405 USA*
{linn, mlbm, rick.luce}@lanl.gov

Abstract

As the amount of scientific information available to researchers increases, the challenge of sifting through the information to find what is truly important to their work increases, as well. In this paper we describe ScienceSifter, a tool that addresses this challenge by enabling groups of researchers and channel editors to create and customize information feeds. Using ScienceSifter, users can combine several information feeds, then filter them by keywords to create a focused information feed. They can view the feed in a shared information space in the form of a list, a list with descriptions, or a hyberbolic tree visualization, and they can save items to a shared list. Thus ScienceSifter can reduce the amount of time researchers spend finding and sharing information. It can facilitate shared intellectual activity and activity awareness among the members of the group.

1. Introduction

Emerging e-Science and grid technologies have the potential to transform the nature of collaborative science. The next generation of tools and repositories will eclipse the current generation in size and complexity and, as they evolve, new models of collaboration and computer-supported collaborative work (CSCW) will co-evolve. The brief history of collaboratories already provides evidence that scientists alter how they work together, and their vision of how they can work together, as the enabling technologies mature. For example, scientists involved in the Upper Atmospheric Research Collaboratory (UARC) in the 1990's realized that the record of a collaborative campaign could be used for retrospective workshops with scientists not involved in the original collaboration [1]. This was an unanticipated result that increased the scale and kind of collaborative work enabled by UARC.

While many collaboration tools are domain-specific, such as collaboratories, some aspects of scientific collaboration transcend individual disciplines, such as the need for shared information. Collaborative efforts, in general, require shared information. Some portion of the information each member of the group understands must be understood by the other members of the group. (Figure 1) The amount of shared information varies depending on the size of the group and the nature of the work. In small interdisciplinary groups, for example, each person may have a unique area of expertise. Without some shared information, however, there can be little or no collaboration.
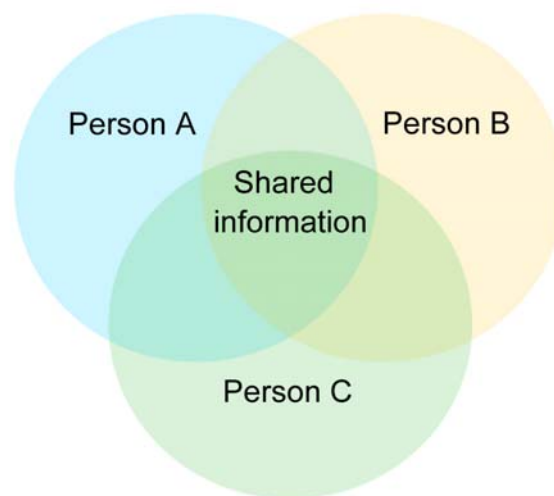

Figure 1. Shared information space

The activities involved in finding and sharing information are time-consuming, from finding

information in the first place, to communicating with others about it and storing it in a manner that makes it possible for others to find it. Maintaining awareness of the information others are finding is also difficult. It is inefficient for everyone in a group to spend large amounts of time engaged in these activities. As the amount of information available to researchers increases, these inefficiencies will increase, as well. It is important to ensure that there is sufficient shared information among members of a research group, without requiring that the researchers, themselves, spend a large portion of their time sharing the information.

One way of facilitating shared information is to create a shared information space, or collaborative webspace. At the Los Alamos National Laboratory (LANL) Research Library, we are prototyping and testing several collaborative webspaces. Weblog-based collaboration portals have been set up for several groups of users, including Research Library staff. Several different models of use have emerged. In one case, the portal is heavily used by a group of five people who post an average of two entries a day. In two other cases, the portals are populated with content by Research Library staff and read by users. Based on researchers' interest in these prototypes and the need for a scalable solution, we were motivated to create the tool described in this paper, ScienceSifter.

Using ScienceSifter, a group of researchers can create a customized information feed by themselves or with the help of a reference librarian or "channel editor". They can select dozens or hundreds of feeds to aggregate into one feed, then use a set of keywords to filter the feed. They can sift through the filtered feed in a collaborative webspace by viewing a list, a list with descriptions, or a hyperbolic tree visualization. They can also save important items.

ScienceSifter utilizes a set of standards and technologies commonly referred to as RSS: Really Simple Syndication, Rich Site Summary, or RDF (Resource Description Framework) Site Summary. Like other RSS feeds, ScienceSifter feeds in the form of a list can be viewed using an RSS reader on a personal computer or hand-held device, as well as in the collaborative webspace described above.

Our work in designing collaborative webspaces is informed by research in activity awareness, an active area of research in computer-supported collaborative work (CSCW). Activity awareness pertains to the kinds of high-level understandings without which groups cannot truly collaborate [2, 3]. These shared understandings are often peripheral to the day-to-day tasks involved in collaborative work and yet they frame and inform the overall direction of projects.

In the long term, collaboration tools will need to meet new challenges that emerge and co-evolve with e-Science technologies and practices. In the near term, certain factors are obvious and must influence design. Our approach to designing collaborative information spaces is influenced by these factors.

2. Design factors

2.1. Large-scale digital object repositories

As e-Science evolves, digital libraries will grow in size and complexity. A scalable architecture for digital object repositories (aDORe) has been developed at the LANL Research Library [4]. An important aspect of these repositories is that they will contain not only more data, but more kinds of data. [5]. Digital library objects will not be limited to published scientific literature, but will also include data sets, the software used to analyze data sets, registries of sensors and instruments, researchers' lab notebooks, pre-prints of upcoming publications, videos of conferences and meetings, and news from journal publishers and scientific societies. Digital library standards have evolved to accommodate these various kinds of data [6, 7].

Once these repositories have been built, the amount of information scientists and engineers can access will increase in size and scope. The LANL Research Library databases currently contain approximately 100 million metadata objects and this number will continue to increase. In order to facilitate collaboration among scientists from different institutions, all participants in the collaboration will need access to these resources. The aDORe repository architecture anticipates and accounts for this by associating every object with properties pertaining to digital rights management. Many policy issues need to be resolved, however, before scientists from different institutions will have universal access to shared information. Researchers at the LANL Research Library have been instrumental in efforts to facilitate this, such as the Open Archives Initiative [8, 9, 10, 11, 12].

2.2. Easy access to information

The point of ingesting more types of information into repositories is not merely to store more information; it is to make more information available to scientists. At the same time that digital object repositories need to grow to include more objects and more varied objects, then, access to the information in these repositories needs to become easier. In order to alleviate information overload, information needs to be

filtered so that only the most important information is provided to users.

The scale of e-Science repositories raises questions about what kinds of mechanisms for accessing and filtering information will be most useful for scientists: search and retrieval, browse, alerts, recommendations, information feeds? While all of these are useful in certain contexts, information feeds may have the most potential, yet they are underutilized: fewer than 2% of adult Internet users currently subscribe to information feeds [13].

In the context of collaborative research groups, information feeds can meet the needs of users far better than traditional search and browse interfaces. Consider a typical search scenario. In order to find information, an individual user decides where to search and how to search: for example, whether to search the Web or a specific database, whether to use basic or advanced search, and which keywords are likely to yield the desired results. The results of the search are displayed to the individual user, who decides whether to download or save them.

Such an individual-centered search and retrieval scenario includes obvious inefficiencies. The scope of the search is limited to a particular individual's knowledge of information resources; the search strategy is limited to the user's knowledge of keywords; and the results of the search are limited to the individual user's computer or account. While the current generation of search tools makes it possible for users to receive vast amounts of information at their desktop, then, it does not support collaboration as well as focused information feeds delivered to a collaborative webspace.

2.3. Timely access to information

Making access to information easy is of little use unless the information is timely. Information feeds refresh constantly and can provide updates to a collaborative webspace 24/7/365. The information in the feeds needs to be up-to-date, as well, or the point of delivering information using this mechanism is lost.

Feeds based on abstracting and indexing databases are not timely, because there is a time lag between the time a paper is published and the time the bibliographic record is accessible in a database. Locally archiving journals makes it possible to provide users with information in a timely manner, soon after it is published and weeks or months before it would be available from a database.

The emerging generation of repositories will index far more than published works. There is a logical flow to scientific activity: data sets exist prior to the papers written about them, and preprints exist prior to

published papers. As more kinds of information become available in repositories, it will be possible to deliver some of this information before it is published. Mechanisms for delivering this information to users in a timely manner will be essential.

2.4. Access to shared information

While it is important for the members of a research group to have access to the same information, it is not efficient for the members of a group to search for the same information individually, duplicating efforts by searching the same resources and using the same keywords. Nor is it efficient for them to subscribe to information feeds individually using RSS readers or tools such as HubMed, which allow individual users to create feeds based on information in PubMed [14]. These tools make it possible for individual users to increase their efficiency by using information feeds, but collaborative tools can make it possible for an entire research group to increase its efficiency by sharing group feeds.

A group feed is even more important if users are searching for information in different resources, in order to ensure that the information is shared. Collaborative efforts in locating information are bound to be more efficient than individual efforts, given that there are so many resources available: for example, approximately 700,000 entries in science blogs at any given time [15]. Efficiencies in finding and managing shared information will become even more important as the amount of available information increases.

2.5. Increased activity awareness for the group

The kinds of activities scientists engage in, and that they need to be aware of, include reading, writing, and publishing. Research in activity awareness demonstrates that members of a group need to be aware of activities that are shared, as well as activities that are not shared [3]. For example, if one member of the group reads more widely than the other members of the group, his or her list of preferred journals may contain journals other members of the group are not reading. Combining these into a single information feed that is filtered by keyword, and allowing users to browse feeds by keyword, is efficient and ensures two things: (1) that other members of the group will not be forced to browse through articles associated with keywords that are not of interest to them and (2) that they will find articles associated with keywords that are of interest to them, even if the journal in the group feed is not one of their preferred journals.

2.6. Expert assistance in accessing information

It is not efficient for individual scientists, or even for groups of scientists, to create information feeds by themselves when a channel editor can work with them to do so. At LANL, only about one in ten researchers has set up database alerts. By creating group feeds based on these individual alerts, so that other members of the same research group receive the same feed, a channel editor can ensure that more researchers receive more information more easily. Moreover, by communicating with the members of the group to learn their needs and work habits and then creating group feeds based on the group's collective requirements, a channel editor can create more informative feeds for that group than any one of them, or any combination of them, can create by themselves.

3. User scenarios

Several user scenarios informed the design of ScienceSifter. In one scenario, users work with a channel editor to create a customized group feed. In another, less efficient scenario, they work by themselves. In either case, the first step is to select several source feeds to combine into one large feed. The second step is to enter a set of keywords to use in filtering and sifting through the large, aggregated feed.

Figure 2 shows a screen shot of the interface for selecting individual feeds to aggregate. It consists of a list of feeds with checkboxes. Each feed, in this example, is a feed of the current table of contents for a journal. These feeds are created by the LANL Research Library for locally-archived journals so that LANL researchers can obtain feeds of current tables of contents even if the journal publisher does not make them available. Any relevant information feed can be added to the set of source feeds to aggregate.
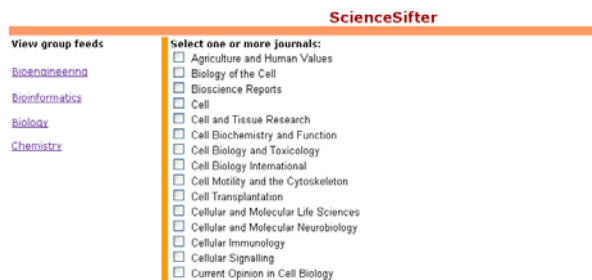
Figure 3 shows a screen shot of the interface for entering keywords. It consists of a list of the feeds already selected for the group feed and a textbox for entering keywords. These keywords are used to filter and sift through the large, aggregated feed.
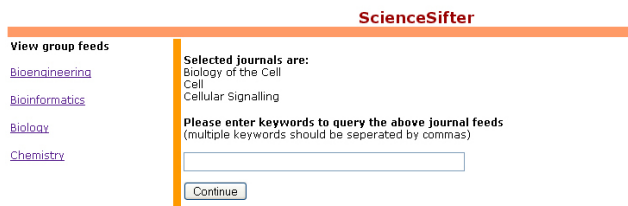

Figure 3. Interface for entering keywords for filtering the aggregate feed

In creating a group feed, channel editors can work with groups of users together at the same time or individually one at a time. The point is to include all of the feeds and keywords of interest to all of the members of the group so that the final feed is an accurate representation of the collective reading of the group.

Once the feed has been created, a web server retrieves and caches the feed daily. Users can view it at any time in a collaborative webspace. In this set of scenarios, they have three options. They can view the entire feed, filtered by all of the keywords, in a hyperbolic tree visualization, or they can select some or all of the keywords and view the feed, filtered by these keywords, in a list of items or in a list of items with descriptions.

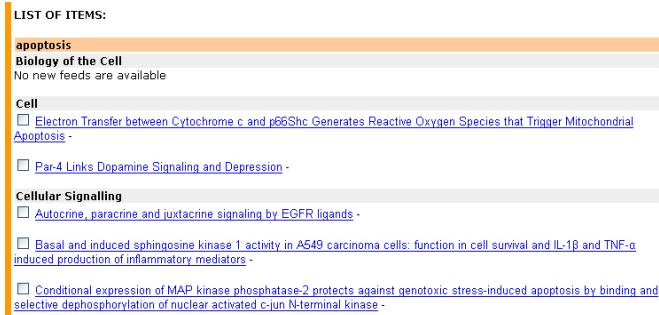As Figure 4 shows, the list of items is concise and scannable.

Figure 4. List of items in the customized feed

Figure 2. Interface for selecting source feeds to aggregate

As Figure 5 shows, the list of items with descriptions provides more information: in the case of journal tables of contents, the abstract of the paper.

Figure 5. List of items with descriptions

Both lists are organized hierarchically, by keyword, then channel or journal, then item or article. While the list of items with descriptions takes longer to read than the list of items, alone, reading the descriptions helps to reduce the number of navigation errors and to increase awareness of the reading activity and research interests of the group as a whole.

As users sift through the lists by keyword and channel, they can select important items to save or bookmark so that they will not be lost when the feed is refreshed. Each item has a checkbox next to it for this purpose.

Figure 6 shows a list of saved or bookmarked items.

Figure 6. List of saved or bookmarked items

Another viewing option, the visualization, provides an overview of all of the keywords, channels, and items for a given group in the form of an interactive hyperbolic tree. This kind of visualization facilitates seeing an overview, zooming in, and viewing details on demand [16]. By clicking on nodes, users can navigate back and forth from one view to the next, sifting through feeds. As they do so, the visualization provides context by means of small iconic representations of the links and nodes.

Figure 7 illustrates three views. In the first view, on the left, all of the keywords associated with a research
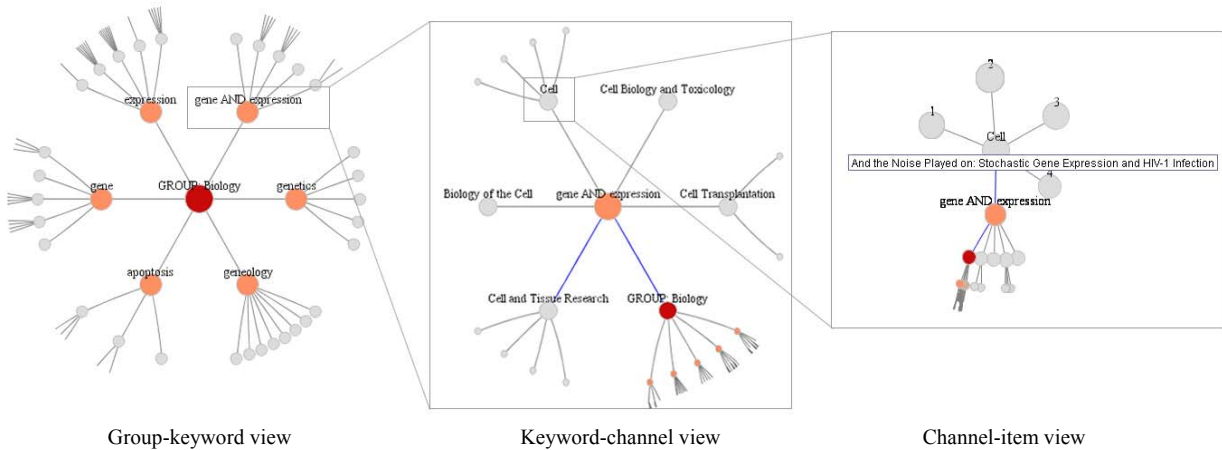

Group-keyword view   Keyword-channel view   Channel-item view

Figure 7. Visualization of customized feed in the form of an interactive hyperbolic tree

group's feed are displayed: for example, gene and expression. In the second view, in the middle, all of the channels associated with the Boolean expression "gene AND expression" are displayed: in this case, feeds for the journals *Biology of the Cell*, *Cell*, *Cell Biology and Toxicology*, *Cell Transplantation*, and *Cell and Tissue Research*. In the third view, on the right, all of the items in the channel for the journal *Cell* are displayed. In this particular case, the channels are journal feeds and the items are journal articles but, of course, many other kinds of feeds can be used with ScienceSifter.

This visualization is based on prior work by the second author, called OncoSifter, which facilitates browsing for information about cancer. In a study of OncoSifter, given a choice of a list or a visualization, every user preferred to see an overview of feeds by using a hyperbolic tree [17].

In addition to viewing ScienceSifter feeds for their own group, users can view the feeds for any other group. This facilitates activity awareness across multiple groups and increases the potential for interdisciplinary research, an important consideration at LANL.

In another scenario supported by ScienceSifter, channel editors and members of a group can modify a group feed at any time by selecting more source feeds or by entering more keywords. This makes it possible for them to continually refine and focus the feed.

4. System architecture

4.1. Overview

ScienceSifter utilizes Urchin, an open source RSS aggregator created by the Nature Publishing Group [18]. Urchin provides a command-line interface for adding, removing, reloading, and aggregating source feeds. It provides a graphical user interface (GUI) for creating a query for filtering source and aggregate feeds. The query can consist of a simple keyword query, a regular expression query, or a Boolean query. Through the GUI interface, users can convert RSS feeds into other formats and vice versa. Feeds are displayed as a list of items without descriptions.

Urchin makes it possible to create custom feeds, which is the motivation for using it as the foundation for ScienceSifter. It is not feasible for our users and channel editors to use the command-line interface, however, which is one of the motivations for creating ScienceSifter. Another motivation for creating ScienceSifter is that the Urchin framework does not support the range of user scenarios we require: groups of users working with channel editors or by

themselves; creating and refining information feeds; viewing information feeds in a shared information space in the form of a list, a list with descriptions, or a hyperbolic tree visualization; and saving or bookmarking items in this shared information space. ScienceSifter utilizes the existing Urchin framework to create information feeds, then, and communicates with Urchin to access information feeds once they have been created, but it provides features that support a wider range of user scenarios. These features render Urchin useful to groups of users who would not otherwise be able to use it and benefit from it.

As Figure 8 illustrates, the ScienceSifter architecture consists of two main components. The configuration component handles the task of configuring feeds for a particular research group. The user interface component supports access to information and the bookmarking feature and displays feeds in the three forms discussed: a list of items, a list of items and descriptions, and a hyperbolic tree visualization. Both components communicate with the Urchin framework.

4.2. Configuration component

This component is used by channel editors or users for configuring group feeds. Three primary purposes are served by this component:

1) Set up feeds for a new research group: When channel editors or users create feeds for a research group, information about group name, journal feeds, and keywords is captured and stored within the aggregate, output, and output_filter tables of the Urchin database.

2) Add, delete, or modify group feeds: These involve SQL queries that manipulate the contents within the output and output_filter tables of Urchin.

3) Reassign feeds to a new group name: If the name of a group changes, the new name replaces the old one with a simple update query for the aggregate table.

4.3. User interface component

This component is used by channel editors or users to check feeds. Users can check feeds for their own group or for other groups at LANL. Reference to the group name is stored in order to access the respective group's keywords and associated channels using a combination of the channel_aggregate, output and output_filter tables from the Urchin database.

This component also handles the group's set of keywords. It displays feeds organized by keyword in a visualization or a list, with or without descriptions.
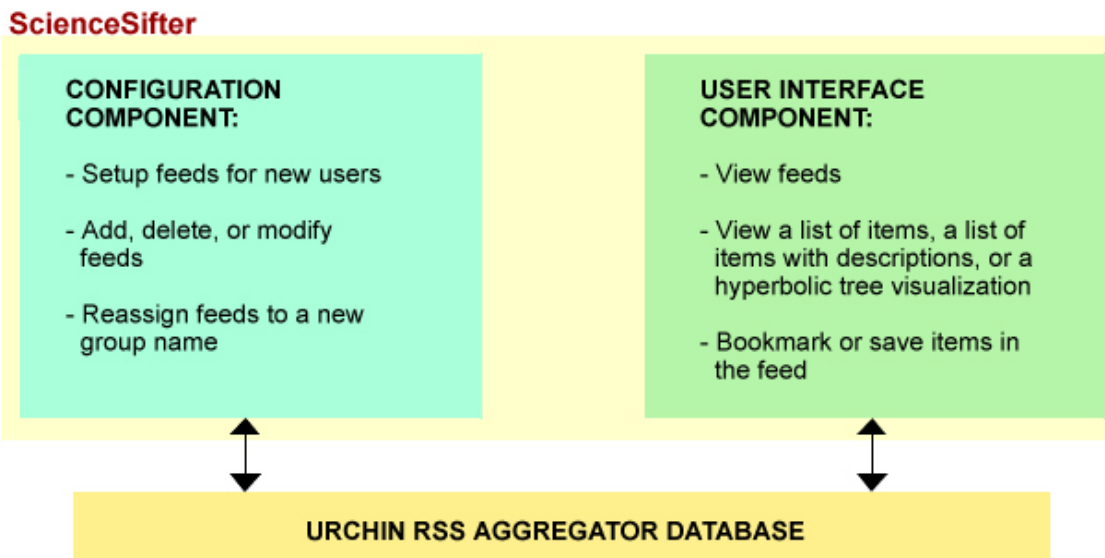
Figure 8. System architecture

A combination of the output, output_filter and channels tables is used to access the latest feeds for one or more keywords. The additional step of constructing an XML file is required for a hyperbolic tree visualization.

A bookmark feature is included in the user interface component to facilitate storage and access of interesting items or articles. All of the bookmarks are stored within a text file with corresponding references within the item table of Urchin. When a member of the group uses the bookmark feature, the corresponding file is read. SQL queries are used to sift through the item table. Only the matching references are captured and displayed as the particular group's bookmarked items.

5. Future work

In order to keep pace with developments in e-Science tools and repositories, ScienceSifter will continue to evolve. As we explore the development of new ScienceSifter functionality, we will contribute to the open source development of Urchin.

We plan to add a feature that will allow users and channel editors to add source feeds to the Urchin database. These feeds will then appear on the interface that users and channel editors use when they select individual feeds to aggregate for a particular group of researchers.

We will be adding more kinds of information feeds. In order to provide LANL scientists with feeds from all journals, not just the ones we archive locally, some of these feeds will come from external journal publishers. Other feeds will come from scientific societies and news organizations. Some feeds will consist of scientists' own weblogs. Two sets of feeds will be created by the Research Library: feeds of group alerts from locally archived bibliographic records and feeds of group bibliometrics, such as recent publications by members of the group and citation counts, from locally archived citation data.

Journal content for locally archived journals will be stored in the digital object repository. Currently, the metadata for locally archived journals is formatted in XML and stored in flat files. Indexes are built with Lucene from the XML files. The feeds are constructed by scraping the XML files and extracting the most recent issue of each journal. This is an I/O-intensive task that will be replaced by another process once journal content is stored in the repository.

We will be presenting more kinds of visualizations in order to facilitate resource discovery. Scientists will be able to browse information feeds by viewing concept maps, taxonomies, co-authorship networks, and scatter plots of quantitative and qualitative data [19], as well as group-keyword-channel-item networks.

We are exploring the possibility of adding another activity awareness feature: sparklines for displaying trends in usage [20]. These will allow users to track

activity by seeing how many times users have clicked on the link to a saved or bookmarked item.

6. Conclusion

Human-computer interaction in the context of e-Science will differ both quantitatively and qualitatively from the past: quantitatively in that more sensors, computers, and scientists will be connected and qualitatively in that the practice of science will be altered by this connectivity. This unique integration of human-computer capabilities has the potential to transform the breadth, depth, and pace of scientific discovery.

The amount of information available to scientists will increase, as will the challenge of identifying the truly important information. ScienceSifter offers one solution to the problem, by making it possible for researchers to create a customized information feed with the expertise of a channel editor. It can facilitate easy and timely access to information for a group of researchers and can increase the activity awareness of the group as a whole.

7. Acknowledgements

8. References

[1] G. M. Olson, D. E. Atkins, R. Clauer, T. A. Finholt, F. Jahanian, T. L. Killeen, A. Prakash, and T. Weymouth, "The Upper Atmospheric Research Collaboratory," *Interactions: New Visions of Human-Computer Interaction*, Vol. 3, May-June 1998.

[2] J.M. Carroll, D.C. Neale, P.L. Isenhour, M.B. Rosson, and D.S. McCrickard, "Notification and Awareness: Synchronizing Task-oriented Collaborative Activity," *Int. J. Human-Computer Studies*, Vol. 58, pp. 605-632, 2003.

[3] G. Convertino, D.C. Neale, L. Hobby, J.M. Carroll, and M.B. Rosson, "A Laboratory Method for Studying Activity Awareness," NordiCHI '04, Tampere, Finland.

[4] H. Van de Sompel, J. Bekaert, L. Xiaoming, L. Balakireva, and T. Schwander, "aDORe: A Modular, Standards-based Digital Object Repository," *The Computer Journal*, 2005.

[5] D.E. Atkins, "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, " 2003.

[6] J. Bekaert, P. Hochstenbach, and H. Van de Sompel. "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library," *D-Lib Magazine*, Vol. 9, No. 11, 2003.

[7] J. Bekaert, P. Hochstenbach, L. Balakireva, and H. Van de Sompel. "Using MPEG-21 and NISO OpenURL for the Dynamic Dissemination of Complex Digital Objects in the Los Alamos National Laboratory Digital Library," *D-Lib Magazine*, Vol. 10, No. 2, 2004.

[8] R. E. Luce, "The Open Archives Initiative: Forging a Path toward Interoperable Author Self-archiving Systems," *College & Research Libraries News*, p. 184, March 2000.

[9] H. Van de Sompel, and C. Lagoze, "The Santa Fe Convention of the Open Archives Initiative," *D-Lib Magazine*, Vol. 6, No. 2, 2000.

[10] C. Lagoze and H. Van de Sompel, "The Open Archives Initiative: Building a Low-barrier Interoperability Framework," *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, 2001.

[11] H. Van de Sompel and C. Lagoze, "Notes from the Interoperability Front: A Progress Report from the Open Archives Initiative," *Research and Advanced Technology for Digital Libraries, Proceedings of the 6th European Conference*, ECDL 2002.

[12] C. Lagoze and H. Van de Sompel, "The Making of the Open Archives Initiative Protocol for Metadata Harvesting," *Library Hi Tech*, Vol. 21, No. 2, pp. 118-28, 2003.

[13] C. Li, "RSS 101 for Marketers," *Forrester Research*, July 2005: http://www.forrester.com/Research/Document/Excerpt/0,7211,37422.html.

[14] J.A.P. Yensen, "Editorial: Almost Everything About HubMed," *Online Journal of Nursing Informatics*, Vol. 9, No. 2, 2005: http://www.eaa-knowledge.com/ojni/ni/9_2/hubmed.htm.

[15] D. Secko, "The Power of the Blog," *The Scientist*, Vol. 19, No. 15, 2005: http://www.the-scientist.com/2005/8/1/ 37/1.

[16] B. Shneiderman, *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Addison-Wesley, Reading, MA, 1997.

[17] K. M. Mane and J. Mostafa, "Exploring the Impact of Information Visualization on Medical Information Seeking Using the Web," HCI International 2005, Las Vegas, NV.

[18] T. Hammond, T. Hannay, and B. Lund, "The Role of RSS in Science Publishing: Syndication and Annotation on the Web," *D-Lib Magazine*, Vol. 10, No. 12, 2004.

[19] L. Marks, J.A.T. Hussell, T.M. McMahon, R.E. Luce. "ActiveGraph: A Digital Library Visualization Tool," *International Journal on Digital Libraries*, Vol. 5, pp. 57-69, 2005.

[20] E. Tufte, "Sparklines: Intense, Simple, Word-sized Graphics", May 2004: http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001Eb&topic_id=1&topic=1.