RC25015 (W1006-070) June 18, 2010
Computer Science

IBM Research Report

Information Curators in an Enterprise File-Sharing Service

Michael Muller, David R. Millen, Jonathan Feinberg
IBM Research Division
One Rogers Street
Cambridge, MA 02142
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Information Curators in an Enterprise File-Sharing Service

Michael J. Muller, David R. Millen, and Jonathan Feinberg
IBM Research, Cambridge, MA, 02142 USA
{michael_muller, david_r_millen, jdf} @ us.ibm.com

Abstract. We report on a social-software file-sharing service within a large company. User-created *collections* of files were associated with increased usage of the uploaded files, especially the sharing of files from one employee to another. Employees innovated in the use of the collections features as "information curators," an emergent lead-user role in which one employee creates named, described collections of resource for use by other employees. This role suggests new work practices and new features.

Introduction

File-sharing has been part of work practices in organizations for decades (for review, see Lee, 2003; Voida et al, 2006; Whalen et al., 2008). The advent of social software has begun to affect file-sharing activities (Reynolds et al., 2007; Schwartz, 2007), just as it has in other aspects of work in organizations and enterprises (e.g., Damianos et al., 2006; John and Seligmann, 2006; Millen et al., 2007; Muller et al., 2008; Thom-Santelli, 2008).

This paper examines an emergent behavior and role associated with file-sharing in an enterprise social software context, namely the preparation of collections of documents for use by others. We call the collectors of the documents "information curators" (see Rubel, 2008, for a similar position about bloggers as curators). Information curators are a special case of the more general role of *intermediaries* who help others to find information (Ehrlich & Cash, 1999;

Muller, 1999). Our investigation of the role of curators in enterprise file-sharing is similar to other, emergent roles in organizational social-computing contexts, such as the roles of evangelist, publisher, and community organizer described by Thom-Santelli et al. (2008) in a social-bookmarking service. Indeed, participatory Web2.0 applications tend to favor user appropriation into novel roles and work practices (Muller et al., 2005). These emergent roles and work practices can serve as lead-user descriptions (Franke et al., 2006), helping to anticipate new work practices and the designs and technologies that will be needed to support them (Kujala and Kauppinen, 2004).

This short paper is organized as follows. The next section provides a brief overview of the enterprise file-sharing service, and compares it with published reports of file-sharing and enterprise social software services. We then describe the traffic and contents of the file-sharing service, highlighting the importance of user-created collections in promoting the downloading of files from the service. The next section presents interview results from 22 of the most-active users of the collections features (i.e., users in the "curators" role). We review the work of "curators" against published requirements for file-sharing services, and we end with implications for design of social software for organizational computing.

The Cattail File-Sharing Service

The Cattail file-sharing service was originally designed as an experiment to reduce the volume of email attachments in IBM's email service. A minimum type of functionality was thus the ability for one user to upload a file, and the ability for a second user to download a file. However, in keeping with social-software concepts, its design quickly evolved into a socially-informed venue for sharing files, networking with other users, discovering new information, and constructing aggregates of files (collections) for individual or shared purposes.

In contrast with the current research focus on peer-to-peer file-sharing networks (e.g., Christin et al., 2005; Johnson et al., 2009; Lee, 2003; see also Voida et al., 2006, for a brief survey of internet peer-to-peer systems), Cattail was designed as a single, centralized server, somewhat similar to the UD Dropbox project for a university setting (Schwartz, 2007). Because Cattail runs entirely within a protected corporate intranet, with full authentication for every user, there have been no known issues with inappropriate sharing of copyrighted materials, or with copyright-owners' countermeasures (e.g., Cristin et al., 2005). Because Cattail does not access individual users' own file systems (other than explicit user-initiated uploads and downloads), there have been few issues of personal data becoming visible to unintended audiences (e.g., Johnson et al., 2009).

Cattail thus shares the same type of operational intranet environment as the Apocrita peer-to-peer system (Reynolds et al., 2007), but with a centralized architecture and enhanced social-networking features. When usage data are

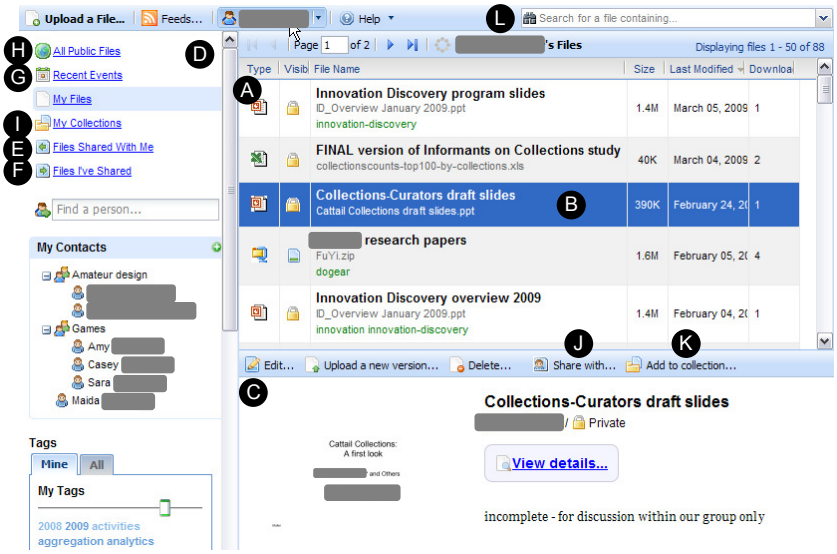

Figure 1. Cattail view of the files related to a particular user. Grey ovals obscure the names of users to protect their privacy.

available from Apocrita and for UD Dropbox, it will be interesting to compare the social interactions associated with each system.

Cattail User Experience

Figure 1 shows a view of a user's own resources in Cattail. The large window on the right (A) contains a list of the user's files. One file has been selected (B), and more information about that file is displayed in the bottom window (C).

The navigation window on the left (D) allows to display different sets of files related to the user, including "My Files" (the current view), files shared *from* other people *to* the user (E), files shared by the user *to* other people (F), recent events in the user's file-sharing network (G), a list of all public files (H), or a list of the collections of files to which the user had access (I). Clicking on the name of a collection displays the files in that collection in a manner similar to Figure 1A.

If the user selected a file (B) and requested to share it (J), the user would be prompted to supply the names of the people to whom the file was to be shared, and an option to send them a message notifying them of the share event (not illustrated). Similarly, if the user selected a file and asked to add it to a collection (K), the user would be prompted to select the collection name from a menu of collections to which s/he had access permissions. The user interface also allows to search all of the files to which the user has access permissions (L). Finally, a detailed view allows the user to add an annotation (or comment) onto the file (not illustrated).

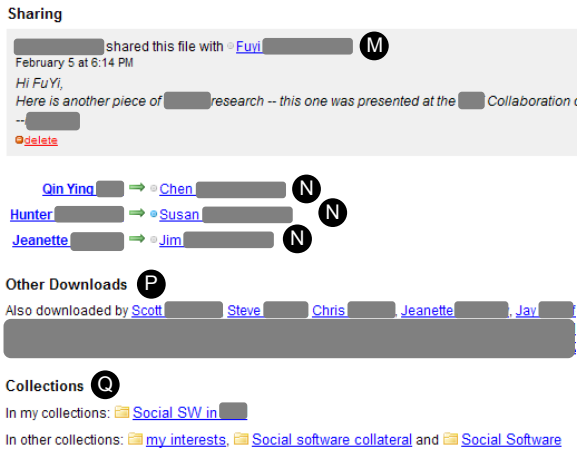

Figure 2. Detail of sharing information about one file. Cattail allows users to track what information was shared, and with whom. Cattail also shows the collections containing the file.

Cattail's functionality reflects in part the analyses of Bellotti (1996), Voida et al. (2006), and Whalen et al. (2008). Voida et al. and Whalen et al. noted user needs to track what had been shared, and with whom – capabilities that are available in the sharing information about each file (Figure 2). Voida et al. also noted user needs to be able to notify others of new or updated material. Cattail records sharing acts by the uploader of the file, including notification messages (M); sharing acts by other users (N); a complete list of all users who have downloaded the file (P); and a list of all collections in which the file has been included (Q); these features come close to the file-history that was advocated by Whalen et al. Bellotti (1996) and Voida et al. (2006) also described a need to be able to specify a group of users, and then to share explicitly with that group; Cattail permits each collection to have a list of members (not illustrated), thus supporting this functionality. Finally, Whalen et al. advocated providing an artifact-based view and a person-based view of file-sharing. Figure 2 illustrates a file-based view, and each person's name serves as a link to a person-based view (similar to Figure 1).

Experiences with File-Sharing in Cattail

The dataset for the quantitative analyses comprised all of the Cattail data records from its introduction on 17 May, 2007 until 10 December 2008. During that time, 15934 employees uploaded a total of 120288 files, which were accessed by a total of 85707 employees (including the 15934 uploaders), who collectively performed 728509 download events. Employees spanned 81 countries, and were employed in a diversity of organizations including product development, sales, marketing, planning, internal operations, and research. Table 1 provides a high-level summary of system usage during the study period.

Resources		Principal Actions		Users in roles	
Files	120288	Upload	120288	Total	88270
Collections	12461	Download	728509	Uploaders	15934
Annotations	8828	Share	107341	Downloaders	85707
		Collect	79823	Sharers	12584
		Annotate	9494	Collectors	5444
				Annotators	3884

Table 1. Cattail resources, actions, and user roles from 17 May 2007 through 10 December 2008.

We were initially interested to predict which files would be downloaded – i.e., Cattail exists to serve files to users through the download operation. While distinct download operations occurred 728509 times, only 89956 unique files (just under 75% of all files) were downloaded. Therefore, despite the purpose of Cattail to share files, more than 25% of the stored files were never downloaded.

Through a multiple regression analysis, we found a higher number of downloads for files that had been shared ($F_{(1,127287)}$=7696, p<.001), collected ($F_{(1,127287)}$=5501, p<.001), and/or annotated ($F_{(1,127287)}$=5104, p<.001).

It seemed obvious to us that sharing a file should increase the number of downloads of that file. We therefore did not focus on the sharing operation.

The work practice of collecting files was more interesting to us. As shown in Table 1, 5444 users created 12461 collections of files through 79823 distinct operations of adding a file to a collection. A total of 60476 unique files (50%) appeared in at least one collection (ten times the number of files with annotations). The phenomenon of collecting affected half of all files. We also found that collections were only marginally associated with refinding of one's own files, ($F_{(1,127287)}$=4.6, p<.05), while they were highly associated with downloading files that had been uploaded by others ($F_{(1,127287)}$=4910, p<.001).

Qualitative Exploration of Collections

To gain insight into the phenomena associated with collections, we undertook an interview study. We identified the 100 most-frequent users of the make-collection feature, including employees from 22 of the 81 countries in the initial dataset of 15934 users. We selected 22 people to interview, using heuristics of trying to maximize diversity in number of countries (16), and to balance women (36%) and men (64%). Informants held a diverse array of job titles, including sales, consulting, business operations, solutions architecture, product management, internal communications, and design.

To reduce costs, and to accommodate large differences in timezones, interviews were conducted via one-hour instant messaging sessions. Interviews were semi-structured, guided by the following topics: motivation for creating collections; users of the informant's collections; presentation/interpretation of the

informant's collections to their users; use of other people's collections. Interview results were coded by a single analyst, through five coding iterations. For this initial analysis, we focused largely on the answers to the above questions.

Informants used collections for individual and shared work: *"easier for myself and others to find the content again… a more active way of sharing"* (Informant 11, internal communications, Austria); *"a knowledge package…. to summarize and to create mixed knowledges [with others]"* (I8, consultant, Turkey); and *"group them in a neat bundle to shove it at people"* (I9, project management, UK) as *"a single focal point of entry"* (I13, enablement, Spain).

Many informants had a semi-structured approach to using collections: *"1. organizing files… 2. finding files I use most often (either my own or those of the people i [work with]… 3. sharing files…"* (I22, sales, USA). In some cases, the collection itself was highly structured: *"collection will contain a master report, and an updated report for every member of my department…. 31 people are able to download reports. 20 people are able to update that report…. save it on cattail so I open a new collection every week…"* (I2, supply chain specialist, Mexico).

Several informants reported using multiple collections for multiple audiences: *"organize by clients or projects… by [human capital management] topic (i.e., workforce planning, hr Business Intelligence) or by project (i.e: customer XX project definition)…"* I17, consultant, Italy); *"Rather than share each [article], I just shared the collection… [project1] internal initiative… [project2] Community calls… [project3] share free ebooks from industry…"* (I20, designer, Canada); *"my team… my boss and his team… virtual development teams… brand executives and their orgs… [world-wide] enablement folks…"* (I22, sales, USA).

Information Curators

Informants created collections with specific intentions for their use, and detailed concern for their audiences: *"regular collections with manually selected / curated resources…. trying to help people (and myself!) make sense of the files that are available…. putting together a collection and deciding what goes into it… and if they are different from the ones I've seen before then I add them to my collection…"* (I15, enterprise 2.0 evangelist, Canada); *"a kind of editor, you share you own and other useful info via collections"* (I18, sales, Finland); *"put some structure around the content I collect/create around my topic… what is good for me is good for my readers ☺"* (I19, product manager, France).

People in this emergent role were concerned to describe or frame their collections for discovery and use by others: *"very short descriptions… the intent of the collection – so I can keep the collection name really short!"* (I9, project management, UK); *"sometimes I used the [descriptive field] to link to other related content [cross] reference"* (I19, product manager, France); and *"i asked everyone to use the naming convention, and I enforced it"* (I22, sales, USA).

These curated or edited collections were intended for both current and future use: *"audit [can] go in on a monthly basis so that they can test to see if the necessary billing approvals exist."* (I4, business operations, UK); *"an asset for the future opportunities about the client"* (I8, consultant, Turkey) and *"It's a fail safe if I was knocked down by a bus!"* (I4, business operations, UK).

Conclusions: Implications for Design

We have shown through quantitative analyses that collections are strongly associated with the use of uploaded files. Collections are particularly important in promoting use by other users – i.e., downloading by a user who had not herself uploaded that file. Collections appear to be key aspects of making a file-sharing service work in a social software environment.

Qualitative analyses illustrated how employees have adopted and adapted collections into their work practices with existing teams and work domains. Qualitative analyses also showed an emergent new role, that of the "information curator," who prepares assemblies of materials for known audiences, and who presents or interprets those materials to those audiences. Curators' collections have lasting value to their audiences and, potentially, to their organizations.

Curators may provide a view into the future of file-sharing, similarly to other "lead user" roles that have helped to define new practices, new opportunities, and new features (Franke et al., 2006; Kujala and Kauppinen, 2004; Thom-Santelli et al., 2008b). Unlike the blogging "digital curators" proposed by Rubel (2008) and others, the information curators in our study often knew their audiences, and collected files to match specific audience needs. Curators have used the social attributes of Cattail to address some of the needs outlined in previous research, such as the ability to share easily with a known audience, to create and use views based on both artifacts (and now *collections* of artifacts) and on persons (Bellotti, 1996; Voida et al., 2006; Whalen et al., 2008).

However, informants also noted gaps in the functionality. While a person-centered view is useful, there are no *group*-centered views, e.g. of the downloads or other actions of the *audience* of a collection. Thus, audience analysis and audience development remain major challenges (see also Thom-Santelli et al, 2008). While a historical view of the actions related to a *single file* is valuable, there is no means for summarizing the history of a *collection* of files. Curators must work hard to understand if their collections are being used, and especially if each collection is providing value *as a collection*. While tagging and annotating are available to clarify the meaning and significance of individual files, there are no similar capabilities to present, discuss, and co-create the meaning and significance of a collection as an aggregate. We will explore these types of new features, and we will be eager to see whether curators emerge as lead users in other organizational file-sharing projects (Reynolds et al., 2007; Schwartz, 2007).

References

Bellotti, V. (1996). 'What you don't know can hurt you: Privacy in collaborative computing.' *Proc BCS HCI '96*. Springer-Verlag, London, UK, January 1996 241-261.

Christin, N., Weigent, A., & Chuang, J. (2005). 'Content availability, pollution, and poisoning in file sharing peer-to-peer networks.' *Proc EC'05*, ACM Press, Vancouver, BC, Canada, June 2005, 68-77.

Damianos, L., Griffith, J., & Cuomo, D. (2006). 'Onomi: Social Bookmarking on a Corporate Intranet.' Position paper in WWW 2006 Tagging Workshop, Edinburgh, Scotland, UK, May 2006, http://www.semanticmetadata.net/hosted/taggingws-www2006-files/28.pdf.

Ehrlich, K., & Cash, D. (1999). 'The invisible world of intermediaries: A cautionary tale,' *Computer Supported Cooperative Work*, vol. 8 no.1-2, p.147-167.

Franke, N., von Hippel, E., & Schreier, M. (2006). 'Finding commercially attractive user innovations: A test of lead-user theory.' *Jour. Prod. Innov. Mgmt*., vol. 23, no. 4, 301-315.

John, A., & Seligmann, D. (2006). 'Collaborative tagging and expertise in the enterprise.' Position paper in WWW 2006 Tagging Workshop, Edinburgh, Scotland, UK, May 2006, http://www.semanticmetadata.net/hosted/taggingws-www2006-files/26.pdf

Johnson, M.E., McGuire, D., & Willey, N.D. (2009). 'Why file sharing networks are dangerous?' *Communications of the ACM*, vol. 52, no. 2, February 2009, 134-138.

Kujala, S., & Kauppinen, M. (2004). 'Identifying and selecting users for user-centered design.' *Proc NordCHI '04*, ACM Press, Tampere, Finland, October, 2004, 297-303.

Lee, J. (2003). 'An end-user perspective on file-sharing systems.' *Communications of the ACM*, vol. 46, no. 2, February 2003, 49-53.

Millen, D.R., Yang, M., Whittaker, S., & Feinberg, J. (2007). 'Social bookmarking and exploratory search.' *Proc ECSCW 2007*, Springer-Verlag, Limerick, Ireland, Sep. 2007, 21-40.

Muller, M.J. (1999). 'Invisible work of telephone operators: An ethnocritical analysis.' *Computer-Supported Cooperative Work*, vol. 8, no. 1-2, 31-61.

Muller, M.J., Geyer, W., Brownholtz, B., Dugan, C., Millen, D.R., and Wilcox, E. (2007). 'Tag-based metonymic search in an activity-centric aggregation service.' *Proc ECSCW 2007*, Springer-Verlag, Limerick, Ireland, September 2007, 179-198.

Muller, M.J., Minassian, S.O., Geyer, W., Millen, D.R., Brownholtz, E., & Wilcox, E. (2005). 'Studying appropriation in activity-centric collaboration.' Position paper at ECSCW 2005 workshop, *Supporting appropriation work*.

Reynolds, J.J., McLeod, R., & Mahmoud, Q.H. (2007). 'Apocrita: A distributed peer-to-peer file sharing system for intranets.' *Proc ACMSE'07*, ACM Press, Winston-Salem, NC, USA, March 2007, 174-178.

Rubel, S. (2008). 'The digital curator in your future.' Micropersuasion http://www.micropersuasion.com/2008/02/the-digital-cur.html, 6 Feb. 2008.

Schwartz, A. (2007). 'UD Dropbox 2.0: Collaboration magic.' *Proc SIGUCCS'07*, ACM Press, Orlando, FL, USA, October 2007, 305-309.

Thom-Santelli, J., Muller, M., & Millen, D.R (2008). 'Social tagging roles: Publishers, evangelists, leaders.' *Proc CHI 2008*, ACM Press, Florence, IT, April 2008, 1041-1044.

Voida, S., Edward, W.K., Newman, M.W., Grinter, R.E., & Duchenault, N. (2006). 'Share and share alike: Exploring the user interface affordances of file sharing.' *Proc CHI 2006*, ACM Press, Montréal, QU, Canada, April 2006, 221-230.

Whalen, T., Toms, E.G., & Blustein, J. (2008). 'Information displays for managing shared files.' *Proc CHIMIT'08*, ACM Press, San Diego, CA, USA, November 2008.