# Reuse in the Wild: An Empirical and Ethnographic Study of Organizational Content Reuse

**Yelena Mejova[1], Klaar De Schepper[2], Lawrence Bergman[3], Jie Lu[3]**

[1] University of Iowa, Iowa City, IA USA {yelena-mejova@uiowa.edu}
[2] Columbia University, New York City, NY USA {kld2116@columbia.edu}
[3] IBM T.J. Watson Research Center, 19 Skyline Dr, Hawthorne, NY USA
{bergmanl@us.ibm.com, jielu@us.ibm.com}

## ABSTRACT

We present a large-scale study of content reuse networks in a large and highly hierarchical organization. In our study, we combine analysis of a collection of presentations produced by employees with interviews conducted throughout the organization and a survey to study presentation content reuse. Study results show a variety of information needs and behaviors related to content reuse as well as a need for a personalized and socially-integrated networking tool for enabling easy access to previously generated presentation material. In this paper we describe our findings and outline a set of requirements for an effective content reuse facility.

### Author Keywords

Content reuse, Social network, Data diffusion, User study

### ACM Classification Keywords

H.1.2 Human Factors, H.4.1 Office Automation, I.7.1 Document Management

### General Terms

Human Factors, Measurement

## INTRODUCTION

Creating and presenting slideshows using software such as Microsoft PowerPoint[1] or OpenOffice Impress[2] is a common activity in the workplace, particularly among knowledge workers. Almost a decade ago, Microsoft estimated 30 million PowerPoint presentations were produced a day [1], a figure that has most certainly increased. A great deal of time is spent producing these presentations.

One way of alleviating this time drain is content reuse. Within our organization, we have observed that reuse of content from existing presentations is common in creating new presentations, especially for time-critical tasks. However, current presentation tools provide little or no support for gathering and sharing materials, leading to frustration and loss of productivity.

These observations raise a number of important questions: How widespread is presentation content reuse within a large-scale and highly hierarchical organization? What are the characteristics of reuse, and what factors facilitate reuse? What are the barriers to reuse? The answers to these questions will inform the development of tools designed to improve the productivity of knowledge workers, through facilitating content reuse and sharing.

To help answer the above questions, we postulated several hypotheses about content reuse (Table 1), and employed both quantitative and qualitative analysis to validate these hypotheses. Our study was conducted in IBM, a global organization of more than 400,000 employees with tens of layers of hierarchical job role structure. We first performed a thorough data analysis of a collection of presentations from an internal IBM file repository, CatTail [2]. Second, we surveyed the authors of these presentations to better understand their reuse habits. Finally, we conducted a series of interviews with IBM employees across the company to study the experiences of both manager and non-manager presentation creators in different IBM divisions.

| |
|---|
| *The need for reuse*<br>**H 1.1** Reuse is wide-spread throughout the organization.<br>**H 1.2** People want to reuse.<br>**H 1.3** There are barriers to reuse. |
| *Reuse characteristics*<br>**H 2.1** Portions of presentations that are difficult to generate (images, charts, graphs) are reused often.<br>**H 2.2** People who know each other (who are close in the social network) reuse each other's materials more often than those who are far.<br>**H 2.3** People from different parts of the organization have different reuse characteristics. |

**Table 1. Hypotheses about content reuse**


*CHI 2011,* May 7–12, 2011, Vancouver, BC, Canada.


[1] http://office.microsoft.com/en-us/powerpoint/
[2] http://www.openoffice.org/product/impress.html

Our study reports on characteristics of content reuse in the organization, with a focus on social factors relevant to sharing, followed by a set of guidelines for developing an effective content reuse facility in an organizational setting.

## RELATED WORK

One way to help knowledge workers keep track of a growing amount of information is to record document history. The history of a document, sometimes called *provenance* [3], can include a list of people that have used it or performed some operations on it. Content versioning systems keep track of provenance metadata that can be used for content management and search.

When no provenance metadata is maintained, reuse detection is useful. Drucker et al., for example, explored techniques for inferring slide reuse to support management of multiple presentation versions [4]. Besides legitimate reuse, this kind of detection is important for plagiarism detection. COPS [5] is a system designed to detect complete or partial copies of text. Duplicate and near-duplicate detection has been used for data cleaning and integration [6], detecting form letters to regulators [7], and even for helping paper reviewers check if a paper matches too closely with another published paper [8]. Our task is similar, in that we are interested in detecting reuse in a set of presentations, broken into slides, where each slide is treated as a document.

A wide variety of techniques have been proposed for duplication detection, ranging from information retrieval metrics [9, 10] to fingerprinting [8, 11, 12] to locality-sensitive hashing [13]. For smaller datasets, more straightforward algorithms have been used to find similarities between documents. Longest Common Subsequence, which can be efficiently implemented using dynamic programming [14], treats documents as a sequence of lines. Edit Distance (Levenshtein Distance) [15], defined as a minimum number of operations (insertions, substitutions, or deletions) required to convert one string into another, can be used to find differences between documents and merge different versions of the same document. We use these two distance metrics in our work.

Material reuse is closely tied to the sociological notion of *diffusion*, defined as the spread of something within a social system [16]. Information usually flows along close social relations. Thus, spatially proximate people are found to influence each other more than those who are far apart. Also, people tend to interact with others who are similar to themselves [17]. In our research we look at the effects that social factors such as acquaintance, common interests, and organizational proximity have on reuse characteristics.

A study by Jensen [3] which logged activities of 17 knowledge workers is close to ours in spirit. The goal of the study was tracking data provenance and determining the effectiveness of provenance cues for document recall. Our motivation is quite different: by tracking reuse characteristics in personal as well as social spheres, we are concerned with facilitating content reuse within a large organization. We also employ a very different methodology from the Jensen study; rather than directly track actions and artifacts via desktop instrumentation, we infer reuse through mining repository content and repository action logs. We feel this opens the door to much larger-scale reuse studies.

| Name | Division | # Authors | # Pres. | # Slides |
|---|---|---|---|---|
| Research | Research | 115 | 512 | 13670 |
| HR | Human Resources | 111 | 411 | 9022 |
| Sales | Sales & Distribution | 130 | 617 | 17434 |
| Product | Software Product Development | 109 | 520 | 17034 |

**Table 2. Statistics of the groups in the data set**

## DATA ANALYSIS

### Data Set

To examine the extent and nature of content reuse, we extracted a set of PowerPoint documents from CatTail [2], an internal IBM file sharing service. For security reasons, only publicly available presentations were used. The dataset consisted of four groups, each consisting of documents created by authors reporting (directly or indirectly) to a particular high-level manager. For example, the Research group included all documents published by employees who were under the Director of Research in the managerial chain. In general, each group represents a portion of the employees within the organization having similar job roles; for example, the Sales group consists of employees in a single geographic sales and distribution region, while the Product group consists of employees from one of the software development groups within the corporation. Table 2 shows the groups, and for each group, the number of authors publishing to the repository as well as the total number of presentations and slides contained within them.

The author sets of these groups are distinct; the management hierarchies are non-overlapping. Group membership was limited to employees in the United States to avoid dealing with non-English language documents. For each person in the group, all publicly shared PowerPoint presentations were downloaded from CatTail and indexed using Slide Library[3] – a slide management tool developed by IBM Research. Using Apache POI[4], we extracted text and information about images for each slide in the presentations.

### Reuse Detection

We employed several measures for detecting reuse in the collection. These included detecting text and image similarity between slides of different presentations.

#### *Text distance metrics*

We assume it is highly likely that exact duplication of a text segment of significant length indicates reuse. Partial

[3] https://apps.lotuslive.com/
[4] http://poi.apache.org/

duplication may also be an indicator of reuse of a portion of a slide's content. We used two metrics to detect exact and partial text duplication in pair-wise slide comparisons:

- *Edit (Levenshtein) Distance* (ED): the number of insertions, deletions and substitutions needed to transform one slide into another [15].
- *Longest Common Substring Distance* (LCSD): the proportion of longest common substring to the combined length of text on the two slides [14].

These distances were computed using dynamic programming, allowing us to compute the distances between every pair of slides in each group. Both distances were scaled to fall in the range between 0 and 1. The smaller the distances are, the stronger the reuse.

#### *Image distance metric*

As with text duplication, we assumed that image duplication is a strong indicator of reuse. Rather than compute pixel-by-pixel comparison between each image in our collection, we used PowerPoint image ID and image size in pixels as means of identifying duplicated images. The image ID meta-tag persists when the image is copy-pasted from one file to another, although not when images are resized, recaptured, or otherwise altered. Thus, image ID (with image size as a double-check) gives us a conservative indicator of image reuse. Once duplicate images are identified, we compute the *image distance* metric between a pair of slides as one minus the Jaccard similarity [18] between these two slides (proportion of the images two slides have in common to the total number of images in both slides).

#### *Reuse indicators*

Text and image distance metrics are combined to compute *binary* reuse indicators. The *exact reuse indicator* has a value of 1 when two slides have both exact text and image duplication (i.e. a value of zero for all distance metrics) and 0 otherwise.

We also define *partial reuse indicators*. To map from a distance metric to a partial reuse indicator, we need to determine cutoff thresholds on the distance values. We do this based on human annotations of reuse instances.

Specifically, we divided the values of each distance metric into ten "buckets" of width 0.1, and asked a human annotator to label 20-30 pairs of slides from each bucket. Using a web interface to view a pair of slides, the annotator determined if the overlap in content of the two slides was sufficient to label it as "reuse". Using the annotations, we compute a *reuse prediction accuracy* (the number of instances considered "reuse" by the human annotator over all annotated instances) for each of the distance metrics. Not surprisingly, the closer the distance is to 0, the more likely it is to signify reuse, and as it approaches 1, the instances of reuse drop off. Thus, small values of any distance metric are an accurate indicator of reuse.

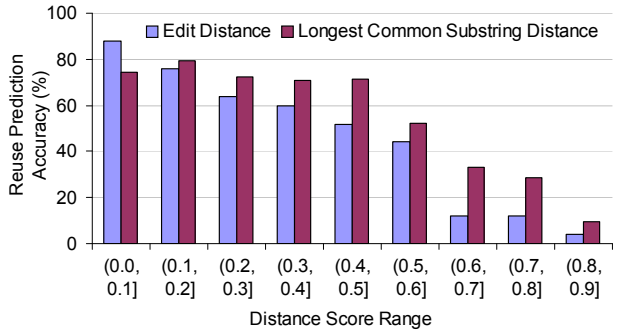



**Figure 1. Reuse prediction accuracy of Edit and Longest Common Substring Distances**

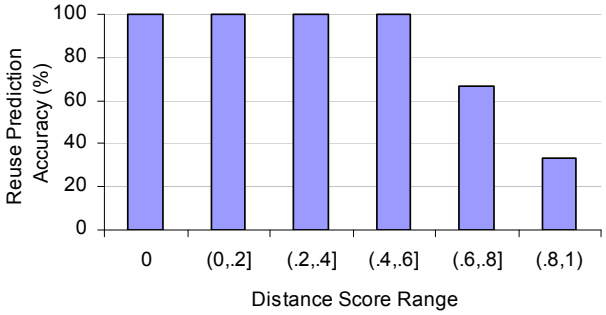



**Figure 2. Reuse prediction accuracy of image distance**

Figure 1 shows the reuse prediction accuracy for the two text distance metrics. Observe that Longest Common Substring Distance (LCSD) remains more accurate in predicting reuse than Edit Distance (ED) as distance score increases (having a reuse prediction accuracy value of 71% compared to 52% for ED at (0.4, 0.5] range). Thus, we selected the value of 0.4 for LCSD to be the upper-bound threshold for a binary *text-based partial reuse indicator*.

Figure 2 shows the reuse prediction accuracy for image distance. Note that its value remains high across a range from 0 to 0.6. Past this point, too few images are duplicated between a pair of slides to be considered reuse. We select the value of 0.6 on image distance to be the upper-bound threshold for an *image-based partial reuse indicator*.

The *partial reuse indicator* has a value of 1 if either the text-based or image-based partial reuse indicator has a value of 1, and is 0 otherwise. The *overall reuse indicator* has a value of 1 when the value of either the exact reuse indicator or the partial reuse indicator is 1.

### Social Network Detection

Once reuse was determined, we related it to the social relationships between authors in the groups of our data set. We used two sources of social networks:

- **SmallBlue** is an IBM social network analysis application that collects outgoing mail and/or chat communications voluntarily shared by IBM employees and combines them with the authorship records of internal blogs, forums, and social bookmark tools [19]. The collection contains records of 30,000+ out of around 400,000 total IBM

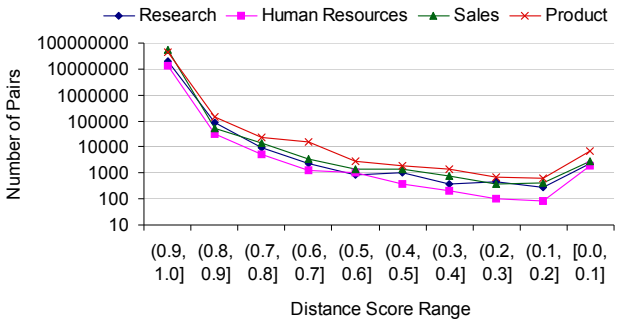



**Figure 3. Distribution of Longest Common Substring Distance scores**

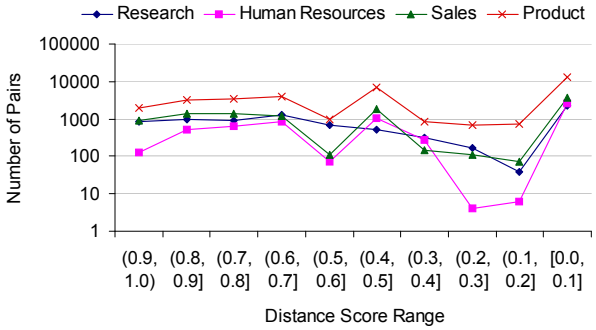



**Figure 4. Distribution of image distance scores**

employees. The network metrics are extrapolated for the employees who are not in the data set [19].

- **SaND** (Social Networks & Discovery) is an IBM Haifa Research project for information discovery and analysis that uses a variety of sources to detect relationships between IBM employees [20]. These sources include relationships via online community sites (blogs, profiles, communities, activities, etc.), wikis, papers, patents, and file sharing using internal repositories. For each pair of employees, SaND provides an overall measure of how well the two people know each other as well as a set of detailed measurements of the relationships between them.

From these two sources, we retrieved the social relationships between the authors within each of the groups. We then used logistic regression to correlate these relationships to the reuse detected in the data set. Table 3 lists the *social relationship variables* we have explored for regression analysis. We will refer to these variables by their labels (specified in parentheses).

### Results

Figure 3 shows the distribution of Longest Common Substring Distance scores. The y-axis is the number of pairs which have a distance score in the indicated range (note the log scale). For this analysis we ignored empty slides, slides which have only stopwords in common (using a short list of most used words in the English language[5]), and "stop-slides" such as those saying "Thank you" or "Questions?"

[5] www.world-english.org

Note that even though the extent of reuse differed for each group, their distance distributions look very similar. Not surprisingly, the distance for a vast majority of the comparisons is at or near 1, meaning that most slides have dissimilar text. But as the distance approaches 0, there is a spike for all groups, indicating a significant amount of exact text reuse, while partial text reuse is comparatively rare. The distribution for Edit Distance looks very much like Figure 3, thus it is omitted for brevity.

The distribution of image distance scores can be seen in Figure 4. This figure omits slide pairs with a distance score of 1 to avoid skewing the display with disproportionately large values along the y-axis. Here, we see a reuse pattern that is substantially more erratic than that observed for text. There is a similar spike at 0, indicating exact image reuse, but there is also quite a bit of reuse in the [0.6, 1] range.

Reuse is fairly prevalent in each of the groups. Figure 5 shows the extent of reuse in each group (the percentage of slides containing reused content). Figure 6 compares the extent of exact vs. partial text reuse. For each group, exact text reuse was more prevalent than partial text reuse, and in the case of Product and HR groups it was almost twice as common.

Figure 7 shows the proportion of reuse by the same author versus reuse by a different author. On average, 14.5% of all slides have reused content, but only 4.4% of all slides have content which has been borrowed from another author. The most active groups were Product and Sales. From Figure 4 we can see that most of the reuse is by authors of their own material, though in the Product group over half of the reused slides are from different authors. Note that the two


**Figure 5. Percentage of slides with reused content**

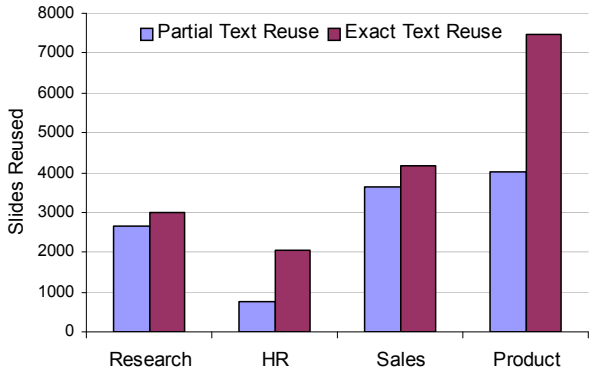



**Figure 6. Number of slide pairs with exact vs. partial text reuse**


**Figure 7. Percentage of reused slides that were reused by the same author vs. by a different author**

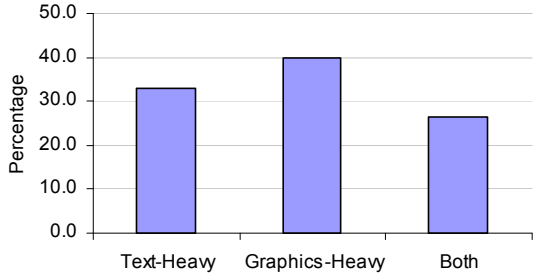



**Figure 8. Percentage of different types of reuse**

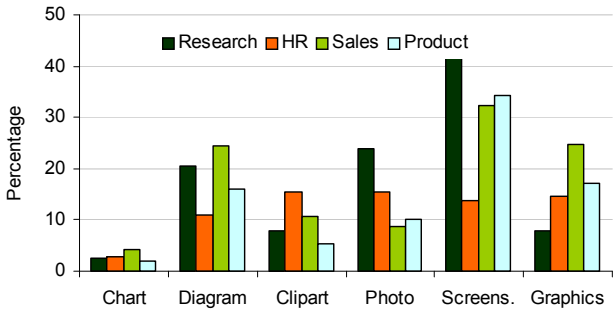
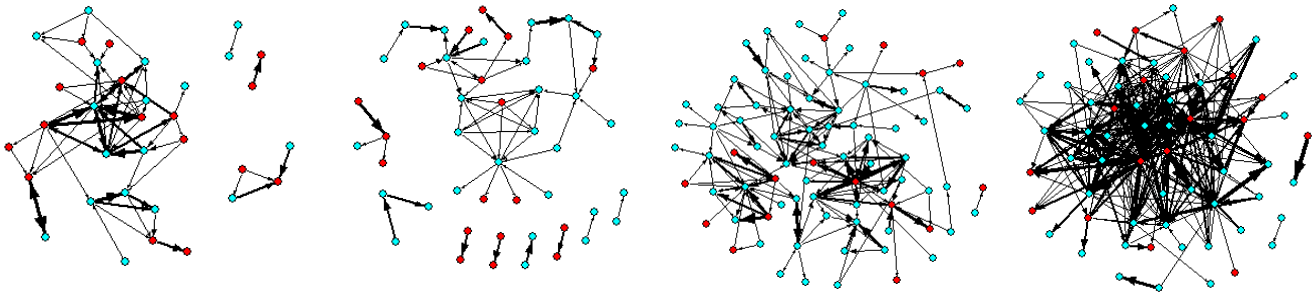



**Figure 9. Percentage of different types of image reuse**

columns in Figure 4 may not add up to 100%, since a slide can be reused by both its author and other authors.

Among the reuse instances (pairs of slides) determined by the overall reuse indicator, we randomly selected a subset of 1228 pairs of slides and annotated them with the types of reuse. Figure 8 shows the extent to which the reuse is text-heavy, graphics-heavy, or both. The "heaviness" was determined by the proportion of textual content to image content based on a visual comparison of the space each took on the slide. Figure 9 further shows the extent of reuse of different types of images. Screen shots, diagrams (a collection of arrows, objects, etc.), and graphics (something created in an image processing program such as Adobe Photoshop or Microsoft Paint) were used the most, followed by photos, clipart, and least of all charts. Note that each group shows differences in the type of images most reused. For example, the Research group particularly favors screenshots, while the Human Resources group uses much less of these compared to other groups. The heaviest user of clip art was the Human Resources group, while the Sales group was the top user of diagrams and graphics.

Using the overall reuse indicator, we constructed directed reuse networks for each group. The direction of reuse was determined using timestamps, assuming that the file with the earlier timestamp is the original. We then examined the topology of the reuse networks using visualization (Figure 10) and by calculating mean clustering coefficients. The direction of the arrows in Fig. 10 signifies the direction of information, from one author to another. Mean clustering coefficients were calculated by dividing the number of instances in which there are reuse ties between people who have a reuse link to the same other person (triangles in the graph) by the number of instances in which three authors are connected by at least two links (connected triples) [21]. These numbers confirmed the differences in reuse characteristics that can be seen in the network visualizations of Fig. 10. Reuse networks in Product (0.36) and Sales (0.30) groups show a greater extent of local clustering than Research (0.14), and HR (0.16), which have a larger number of isolated authors and are sparser. However, when authors who only reuse from themselves and authors who only reuse from one other person are removed [22], these numbers are much more similar for the Product (0.67), Sales (0.61), Research (0.55), and HR (0.51), groups. This means that the density of local clustering of authors who are connected to more than one person is similar between the networks. Mean in and out node degrees, which give an indication of the level of interconnectedness of the entire network, are 5.20 for the Product group, 2.25 for Sales, 0.91 for Research, and 0.95 for HR.



**(a) Research** **(b) HR** **(c) Sales** **(d) Product**

**Figure 10. Reuse network visualization (manager – red, non-manager – blue)**

| |
|---|
| **Blogs** (*g_blogs*) – commenting on each other's blog posts, sharing content, following one another on a blogging site. |
| **Co-authorship** (*g_coauthor*) – co-authoring a paper or a patent, editing the same wiki. |
| **Communities** (*g_communities*) – community co-membership. |
| **Content sharing** (*g_sharing*) – sharing content in a content sharing site. |
| **Commenting** (*g_commenting*) – commenting on each other's files, bookmarking, tagging. |
| **Friending** (*g_friending*) – being "friends" on a social website, profile tagging. |
| **Managerial chain** (*g_mgrchain*) – having the same direct or indirect manager. |
| **From manager** (*from_mgr*) – information reused is from a manager. |
| **To manager** (*to_mgr*) – information is reused by a manager. |
| **Same manager** (*same_mgr_num*) –both the author and the re-user of a content item report to the same immediate manager. |

**Table 3. Social relationship variables**

To verify the significance of social relationships in reuse, we performed a regression analysis. Using logistic regression we determined the odds ratios for each of the social relationship variables (Table 3) and two variables that aggregate all relationships between two people, provided by SmallBlue and SaND respectively. Figure 11 shows the odds ratios and their 95% confidence intervals for the statistically significant variables (statistically insignificant variables are not shown here due to space considerations). As a measure of effect size, odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. For example, for the *same_mgr_num* variable, the value of the dot (4.42) is the ratio of the odds of reusing content if the authors have the same immediate manager relative to the odds of reusing content if they are not under the same immediate manager. The dotted line at 1 signifies the odds ratio where there is no statistically significant association between the variable and reuse.

In Figure 11, the strongest correlation we see is from SaND and *same_mgr_num*, indicating that certain social

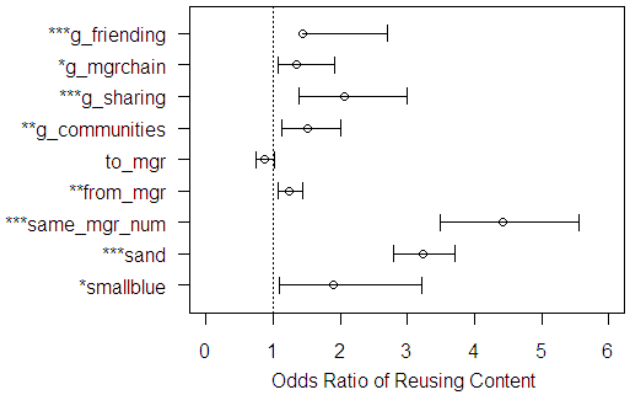



**Figure 11. Odds ratio of social relationship variables (*** p<0.001, ** p< 0.01, * p< 0.05)**

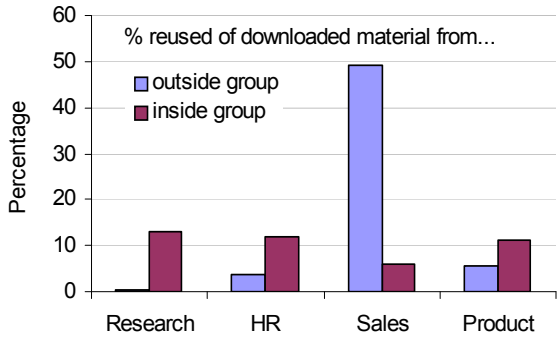



**Figure 12. Percentage of downloaded material being reused**

relationships are indeed important in predicting reuse. There is a positive correlation between *from_mgr* and reuse, meaning that material is more likely to come from a manager than a non-manager, though there is no significant correlation between *to_mgr* and reuse.

Notice that not all relationships have proven to have a significant correlation with reuse. It was especially surprising to find that *g_coauthor* was not a significant variable. Upon further inspection we found that it was the least represented variable in the data set, and thus too sparse for use in statistical analysis.

Finally, using file download records, we looked at the extent of reuse as related to file downloads both within and between groups. For each download, we compared the downloaded document to all documents posted by the downloader with dates more recent than the download. We then calculated the percentage of downloaded material that is reused (see Figure 12). Again, we see a variety of behaviors among the groups. The most drastic is the Sales group, in which probability of reusing outside material when downloaded is much higher than the probability of reusing material from their own group.

## SURVEY

We conducted a survey to gather more information on the reuse habits of the authors whose presentations we analyzed in the previous section.

### Methodology

The survey consisted of ten online questions. Most questions consisted of several parts, and each part had multiple choice answers. For example, when asked how much they reused certain types of presentation content such as images or bullet points, the participants could answer "*never*", "*seldom*", "*sometimes*", "*often*" and "*very often*" for each type. The questions were designed to validate our findings from the data analysis, as well as to clarify questions that could not be answered by examining the presentations. For example, we asked the participants to explain why they reused material from other presentations, or how they went about seeking material from people they did not directly know. The survey was emailed to all 465 authors from our four analysis groups who had publicly available presentations published to CatTail. Eighty-eight authors responded.

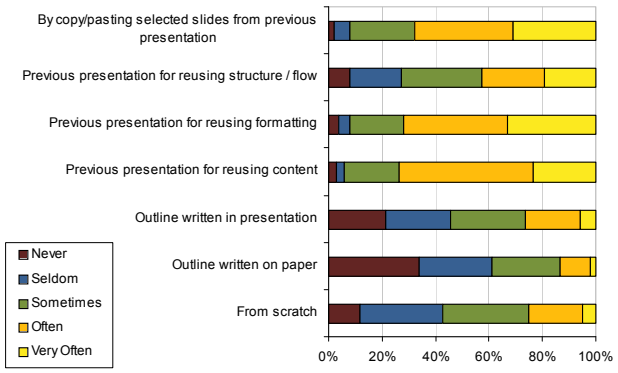



**Figure 13. Question: Select how often you start composing a presentation with the following …**

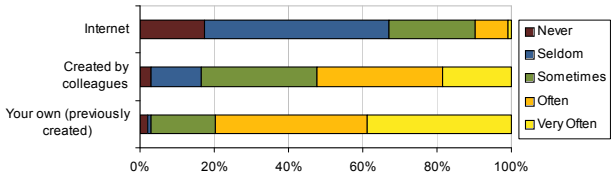



**Figure 14. Question: Where do you get the material for your presentations?**

## Results

The results show the importance of content reuse in making presentations. Figure 13 shows responses to the question "*Select how often you start composing a presentation with the following...*" The three options that did not involve reuse were by far the least popular; "*often*" or "*very often*" were selected by only 26% for outlines written in presentation, 13% for outlines written on paper, and 25% for starting from scratch. On the other hand, 94% said they start with a previous presentation for reusing content at least sometimes. Figure 14 shows responses to the question "*Where do you get the material for your presentations?*" Not surprisingly, most respondents use their own material "*often*" and "*very often*" (79%). However, 83% at least sometimes use materials created by their colleagues. Internet did not prove to be a widely used source of material, with only 10% using it as a source "*often*" or "*very often*".

When asked about external sources for presentation material, the most popular answer was "*people in my direct team*" with 96% selecting "*sometimes*", "*often*" or "*very often*". Two other most popular answers were "*people in my department but not in my direct team*" with 64% and "*information/knowledge repository in my organization*" with 61% selecting at least "*sometimes*". Interestingly, respondents who were managers selected at least "*sometimes*" with about the same frequency for "*people I manage*" and "*people who are my managers*" (89% versus 84%), suggesting that information flows both up and down the managerial hierarchy.

Next, respondents were asked about particular elements of the presentations they created. Images were reused the most, with 62% selecting "*often*" or "*very often*", whereas text paragraphs were reused the least, with 26%. Consistent with the finding from the data analysis, entire slides were used quite a bit, with 85% selecting at least "*sometimes*". Recall that the reuse of an entire slide was shown to be at least as frequent as that of some part of its content (see Figure 6).

The reason most frequently cited for reuse is "*content is time consuming to re-create*", followed by "*to maintain consistency with content previously presented*" and "*content is difficult to recreate*". 79% of respondents selected at least "*sometimes*" for "*you need somebody else's data/content*".

We also analyzed differences in responses between individual groups. When asked which kinds of presentations they produced, with a variety of choices from technical presentations to instruction/teaching, respondents of all groups selected all of the choices (in different proportions). This suggests that there is diversity within each group – Sales people also go to conferences, and Research people also make presentations for potential customers. Reuse characteristics did differ significantly between the groups. The Sales and Product groups reused their presentations and searched for materials on the internet much more compared to Research and HR. The Sales group used information/knowledge repositories quite a bit. The Product group relied on people in their direct team for content the most. Sales people reused materials from people they did not know, whereas Research and HR were much less likely to do so.

Finally, we split the population by managerial status. Managers were slightly more likely to reuse presentation content (84% selected "*often*" or "*very often*") compared to non-managers (71%). They were also much more likely to copy/paste slides from one presentation to another (84% vs. 64%) and use materials created by colleagues (74% vs. 47%).

Of the 88 respondents, not one said he/she has never reused material – the practice is ubiquitous and is an important part of the composition process. The findings from the survey generally agree with those of the data analysis. Although an author's own material is reused most often, colleagues are an important source of content. Authors favored images over text, and often reused whole slides instead of some portion of the content. Differences between the groups were also confirmed. Besides validating the results from the data analysis, we asked people about their motivations for reusing content. The most popular reason was that it was either difficult or time consuming to recreate content.

## INTERVIEWS

To get a better understanding of the needs and wants of presentation creators, particularly with regards to content reuse, we conducted interviews with IBM employees.

### Methodology

We interviewed 24 IBM employees: 16 managers, 8 non-managers. 16 participants (due to ease of recruitment) came from within the company's research division; the other 8, of which 5 were selected from a list of high-volume CatTail users, were from 8 different departments. The interviews, which lasted for 30-60 minutes, were guided by 9 questions as well as examples from three or more sample presentations supplied by the participants. We asked the participants about the kinds of presentations they create, and their processes for finding and reusing materials as well as collaborating with team members.

### Results

All of the 24 participants reuse material when they create slide presentations, both from their own previous presentations and by reusing slides from others. 71% of the participants stated that they create content from scratch only when they cannot find it elsewhere. Participants prefer reusing material as it saves them substantial time; researchers stated that detailed technical descriptions can sometimes take an hour per slide to create from scratch. During the interviews, we noted several prominent patterns of content reuse, e.g.:

- A manager requests previously produced slides from his/her direct reports for inclusion in a status update. The manager then curates and orders the material and adds an introduction, outline, and conclusion for the slides — cited by 58% of the participants.
- The presentation creator needs to give an overview presentation of corporate initiatives. The range of the material needed is beyond what is available in the slide creator's direct professional network. Access to "strangers" and their materials is crucial to the reuse needs in this scenario — cited by 46% of the participants.
- The presentation creator needs to combine pre-sales marketing, sales enablement content, and customer stories, within presentations meant for cross-division sharing by both marketing and sales people. — cited by 13% of the participants.

The reported proportion of reused material from others was especially large for people whose job is to package and present other people's work. For example, the seven participants who create presentations for sales representatives estimated that they create only 10% of the material from scratch. The remaining 90% is created by reusing material, with more than half the reused slides created by others. These participants agreed that it is not only time-efficient but also crucial from a marketing perspective to reuse material. All 16 manager participants collect material from people who report to them for status updates and presentations to peers, and reported a similarly large proportion of content reuse for these types of presentations.

A variety of tools and strategies are used to enable reuse. In order of prevalence, colleagues share presentations as e-mail or instant messaging attachments, links to internal file sharing systems, or via project wikis or other online collaboration environments. 38% of the participants reported using desktop search tools that index presentation content. Several participants search their e-mail database by date, sender, or keywords. Participants agreed that e-mail provides useful provenance metadata, as the subject/body of e-mail messages is likely to include keywords describing the presentation.

Most of the material reused comes from people's direct colleagues. Several participants stated that their network of contacts, and knowledge of who does what in the organization, is crucial for finding and obtaining the right materials. They found it difficult to find company-related material beyond their social network.

Provenance metadata on a slide-level is important across all divisions. 67% of the participants stated that when the origin of reused material cannot be traced, they feel unsure about using the material, lacking assurance of both proper attribution and use rights. They stated that this may keep them from using material, especially for external presentations. 33% of the participants would like to have access to a library of high-quality company-specific clipart. Researchers cited a need for access to publications and sources of data visualizations.

Finally, participants expressed frustration over the lack of integration among the multitude of available knowledge sharing systems. Participants stated a need for better tools that support version control, and side-by-side comparison of slides during iterative collaborative processes.

## DISCUSSION

Confirming our informal observations, one of the main findings of our study is that content reuse is a common practice within the organization. Both the data analysis and the survey revealed a set of diverse reuse characteristics for each of the groups, suggesting the importance of an employee's job role in determining reuse behavior.

Furthermore, we find a close relationship between social ties and the reuse characteristics, confirming sociological studies of information diffusion [16, 17]. People tend to reuse material from people in their social vicinity – "friends" on a social networking website, or their colleagues reporting to the same manager. Unless their job role requires them to communicate with people they do not know (as in sales, for example), people tend to reuse material from people in their social vicinity.

Returning to the hypotheses we have stated in the introduction (Table 1), we are able to draw conclusions about each:

*The need for reuse*

**H 1.1** Reuse is wide-spread throughout the organization. *Confirmed.* Both the data analysis and the qualitative inquiry have showed that reuse is pervasive throughout the company. The data analysis shows that on average, 14.5%

of all slides show at least partial reuse, with 4.4% of slides coming from a different author. The minimum reuse in any of the groups we sampled was 12%. In the survey, 93% of the respondents start with an existing presentation at least sometimes when creating a new presentation, 79% reuse their own materials often or very often, and 83% reuse materials from colleagues at least sometimes. Survey respondents and interview participants stated time savings, difficulty of reproduction, need to maintain consistency, and need for others' data/content as primary reasons for reuse. Furthermore, both the data analysis and the survey have shown the prevalence of whole-slide reuse. Partial reuse does occur (particularly of images), but the entire slide seems to be a natural unit of reuse.

**H 1.2** People want to reuse. *Confirmed*. Interview participants stated that they will avoid creating slides from scratch if they can reuse them, especially ones containing material that is difficult or time-consuming to re-create. Some jobs, such as sales representative, require gathering data from various sources within the company.

**H 1.3** There are barriers to reuse. *Confirmed.* In the course of the interviews, participants cited barriers to content reuse, including: lack of a centralized repository, need for thorough content indexing, insufficient tools for slide material search and preview, and inadequate authorship attribution. Existing software that is meant to solve some of these issues suffers from a lack of centralization and integration with existing presentation creation tools.

*Reuse characteristics*

**H 2.1** Parts that are difficult to generate are reused often (images, charts, graphs). *Confirmed*. People reuse graphics-intensive slides, especially those containing screenshots, diagrams, graphics, and photos. 66.3% of all annotated reuse instances were either image-heavy or both image- and text-heavy. Survey participants indicated that images were reused the most, with 62% selecting "*often*" or "*very often*", whereas text paragraphs were reused the least (26%). The most prominent reason cited for this pattern was the time required to create complex graphics.

**H 2.2** People close in the social network reuse each other's materials more often than those who are far. *Confirmed*. People who know each other or who have similar interests are more likely to share materials. Regression analysis has shown social interactions to be significant in reuse behavior; SmallBlue and SaND social relationship variables have a positive effect on reuse. Authors who are connected through social networking websites (those who "friend" each other), who share resources on public forums, and those in the same communities are significantly more likely to reuse each other's materials. Not surprisingly, people within the same management chain are likely to reuse from each other. Survey results confirm this; 96% use materials from their own team often or very often. Furthermore, our survey results and interviews show that it is difficult to find sought-after resources if they are not available in the author's immediate social network.

**H 2.3** People from different parts of organization have different reuse characteristics. *Confirmed*. People with different job roles reuse materials differently. We have found a wide range of reuse characteristics across corporate divisions. The Product group is the most active in reusing materials within the group. The Sales group heavily reuses materials it downloads from other groups, whereas the Research group is very unlikely to reuse materials from the outside. These behaviors were also clearly seen in the survey responses, where, for example, the Sales group favored information/knowledge repositories while the Product group relied heavily on people in their direct team for content. This suggests that each group has its own needs, and requires special accommodations to facilitate reuse.

The survey has also shown managers to reuse more than non-managers, yet this distinction did not appear in our dataset analysis. From the slide-sharing workflow described by interview participants, we learned that the exchange of information from non-managers to managers is usually not publicly shared or takes place on communication channels other than file sharing services such as CatTail. It is likely that the flow of information up and down the managerial hierarchy takes place via different channels, which shows the strength of using both empirical and ethnographic approaches to capture these trends.

## CONCLUSIONS AND FUTURE WORK

In this study we explore characteristics of content reuse in a large and highly hierarchical organization. This study combines quantitative analysis of a collection of presentations made across our company with qualitative study including a survey and interviews. We have documented the content reuse behavior of several groups of IBM employees, and have shown a diversity of information needs across various job roles. The social network analysis has shown both organizational and personal ties to be important in reuse. These findings are important in the design of a useful tool for facilitating content reuse.

Throughout our interviews we encountered a general dissatisfaction with available content sharing tools. Although slide presentations are an important tool for collecting and sharing knowledge, current systems for sharing slide material do not support effective information management, as little metadata about slide material is maintained. Software tools are needed to help tie the patchwork of knowledge in digital objects more closely to social networks [22, 23].

Based on our findings, we have developed a set of requirements for an effective content reuse facility:

1. It should be "socially-aware", simplifying reuse from people the author knows. This requirement is based on the strong connection between social proximity and

reuse found in the data analysis and interviews. Integration with existing social networking software would provide a centralized repository with improved search ranking, customized permission setting, and team-specific versioning and tagging for each project. Authors who frequently collect materials from particular individuals might have a "favorites" section where the desired materials are readily accessible. Social search outside an author's immediate social group must be provided for tasks involving gathering of information from elsewhere in the organization.

2. Based on interview feedback, a tool should support automatic tracking of the origin of materials and their use. Collection of provenance metadata would simplify authorship attribution, and facilitate collaboration.
3. A tool should be customizable for the author's job role and tasks. The interviews and data analysis have shown a variety of reuse characteristics within different parts of the organization. Various search options should be available: papers for researchers, knowledge repositories for sales representatives, project descriptions for marketing.
4. Based on the preference for reuse of imagery shown in the data analysis, a tool should have good multimedia support. Image search, browsing, previewing and copying should be supported by the appropriate metadata, which may come from the text of the associated slide.
5. The interviews suggest that a tool should support the overall workflow cycle of slide presentation creation and reuse. Slide version comparison, version merging, support for context-specific material recommendations, and interactive placeholders with follow-up contingencies should be provided.

This work suggests several directions for future investigations. We have seen that partial reuse of slides is prevalent, but have not looked in detail at other units of reuse. In examining reuse, we have, for example, seen reuse of groups of related slides. Determining reuse units is a problem for other media types, such as text documents, which do not have a natural unit of reuse such as the slide. Finally, we are seeking to direct the development of a presentation repository system. Our team has developed an initial system, known as SlideLibrary, which supports slide-level sharing and searching. A natural next step is developing a reuse facility that incorporates the above requirements.

## ACKNOWLEDGEMENTS

We wish to thank Julie MacNaught and Danny Yeh for software support in our use of SlideLibrary, and Ravi Konuru for management support. Ching-Yung Lin supplied some of the network analysis and many helpful suggestions.

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## REFERENCES

1. Parker, I. *Absolute PowerPoint: Can a Software Package Edit our Thoughts?* The New Yorker, May 28, 2001.
2. Muller, M., Millen, D.R., and Feinberg, J. *Information curators in an enterprise file-sharing service*. ECSCW '09.
3. Jensen, C., Lonsdale, H., Wynn, E., Cao, J., Slater, M., Dietterich, T. G. *The Life and Times of Files and Information: A Study of Desktop Provenance*. CHI '10.
4. Drucker, S. M., Petshnigg, G., Agrawala, M. *Comparing and Managing Multiple Versions of Slide Presentations.* UIST '06.
5. S. Brin, J. Davis, H. Garcia-Molina. *Copy detection mechanisms for digital documents*. SIGMOD '95.
6. Monge, A. E. *Matching Algorithms within a Duplicate Detection System*. Bulletin of the Technical Committee on Data Engineering 23, 4. 2000.
7. Yang, H., Callan, J. *Near-Duplicate Detection for eRulemaking*. Conf on Digital Government Research 89: 78-86, 2005.
8. Schleimer, S., Wilkerson, D. S., Aiken, A. *Winnowing: Local Algorithms for Document Fingerprinting*. SIGMOD '03.
9. Chen, X., Francia, B., Li, M., Mckinnon, B. *Shared Information and Program Plagiarism Detection*. IEEE Transactions on Information Theory, 50 (7): 1545-1551, 2004.
10. Cilibrasi, R., and Vitanyi, P. *Clustering by compression*. IEEE Transactions on Information Theory, 51(4), 1523-1545, 2005.
11. Broder, A. Z. *On the resemblance and containment of documents*. Proc. of Compression and Complexity of Sequences, 1997.
12. Manber, U. *Finding Similar Files in a Large File System*. USENIX '94.
13. Gionis, A., Indyk, P., Motwani, R. *Similarity Search in High Dimensions via Hashing*. VLDB '99.
14. Hirschberg, D. S. 1975. *A linear space algorithm for computing maximal common subsequences*. Comm. of the ACM. 18, 6: 341–343.
15. Levenshtein, V. I. 1966. *Binary codes capable of correcting deletions, insertions and reversals*. Soviet Physics Doklady.
16. Stang, D., Soule, S. A. *Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills*. Annual Review of Sociology, 24: 265-290, 1998.
17. Rogers, E. M. *Diffusion of Innovations*. New York: The Free Press. 1995.
18. Jaccard, P. *Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines*. Bulletin de la Société Vaudoise des Sciences Naturelles 37: 241-272. 1901.
19. Wen, Z., Lin, C. *On the Quality of Inferring Interests From Social Neighbors*. KDD '10.
20. Ronen, I., Shahar, E., Ur, S., Uziel, E., Yogev, S., Zwerdling, N., Carmel, D., Guy, I., Har'El, N., Ofek-Koifman, S. *Social Networks and Discovery in the Enterprise (SaND)*. SIGIR '09.
21. Watts, D., Strogatz, S. *Collective dynamics of 'small-world' networks*, Nature, vol. 393, no. 6684, pp. 440–442, 1998.
22. Kaiser, M. *Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks.* New J. Phys. 2008.
23. Krebs, V., Holley, J. *Building smart communities through network weaving*. Appalachian Cntr for Econ Networks, 2006.