DriveGroups: Using Group Perspective for Usable Data Sharing in Research Collaborations

JAMES G. FRAZIER, EMILY K. WEINSTEIN, IRIS IZYDORCZAK, and YIFAN WU, Hamilton College, USA
NAZARET CUADROS, UBS, USA
DIPASHREYA A. SUR, Stanford University, USA
SARAH MORRISON-SMITH*, Hamilton College, USA

Sharing data with collaborators is a complicated task that is nonetheless fundamental to academic research. We present the results of two studies investigating data sharing within academic scientific collaborations, as well as a system called DriveGroups designed to facilitate data sharing. First, we observed and interviewed 38 academic researchers engaged in collaborative research about their data sharing practices. We found that these researchers struggle to manage access to data, especially when different types of collaborators (e.g., students, co-principal investigators) require different access settings. In response, we built DriveGroups, a Google add-on designed to alleviate participant challenges with access control, and compared its usability to unmodified Google Drive. DriveGroups allows users to manage file access from two separate perspectives: 1) the traditional file perspective and 2) a role-based group perspective, which simplifies the data sharing process. DriveGroups matched or outperformed unmodified Google Drive in terms of usability, access control, and transparency, and will help scientists advance high-impact academic research.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Open source software**.

Additional Key Words and Phrases: Data sharing; usability; systems; role based access control

ACM Reference Format:
James G. Frazier, Emily K. Weinstein, Iris Izydorczak, Yifan Wu, Nazaret Cuadros, Dipashreya A. Sur, and Sarah Morrison-Smith. 2025. DriveGroups: Using Group Perspective for Usable Data Sharing in Research Collaborations. *Proc. ACM Hum.-Comput. Interact.* 9, 1, Article GROUP13 (January 2025), 28 pages. https://doi.org/10.1145/3701192

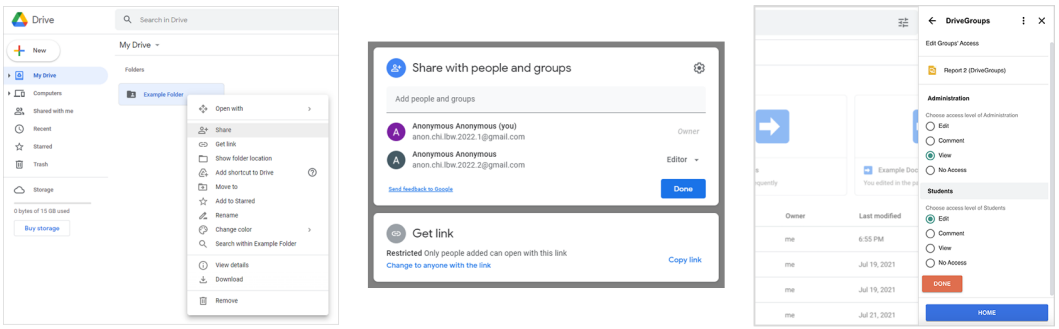
1 Introduction

Generating and sharing large amounts of data enables academic scientific collaboration [1, 39], but it poses significant challenges due to diverse data characteristics and methodologies across disciplines [5, 15, 50]. Managing research data differs notably from other data-sharing practices; it involves ensuring data integrity and reproducibility, enforcing strict quality control for publication readiness, and handling extensive metadata for detailed contextual understanding [26]. Research

*Corresponding author.

Authors' Contact Information: James G. Frazier, jgfrazie@hamilton.edu; Emily K. Weinstein, ekweinst@hamilton.edu; Iris Izydorczak, iizydorc@hamilton.edu; Yifan Wu, ywu2@hamilton.edu, Hamilton College, Clinton, New York, USA; Nazaret Cuadros, nazaret.cuadros@ubs.com, UBS, Weehawken, New Jersey, USA; Dipashreya A. Sur, dipashre@stanford.edu, Stanford University, Stanford, California, USA; Sarah Morrison-Smith, smorriso@hamilton.edu, Hamilton College, Clinton, New York, USA.

ACM 2573-0142/2025/1-ARTGROUP13
https://doi.org/10.1145/3701192



(a) (b) (c)

Fig. 1. Google Drive's sharing interface requires users to navigate a series of steps (a & b) for each individual user and file to grant different levels of access to files in a directory. DriveGroups (c) allows users to create reusable settings for groups of collaborators.

data management also requires adhering to ethical and regulatory standards, especially with human subjects or sensitive data, and addressing requirements of grants for public accessibility and long-term storage. Moreover, collaborations often span multiple disciplines, necessitating interoperable data formats and a grasp of varied terminologies [39].

While numerous bespoke technologies have been created to support publicly sharing finalized data, the confidential nature of unpublished data makes use of these systems for internal data sharing impractical. Furthermore, prior work shows that data sharing is a complex process facing challenges including risks of theft, peer competition, and misinterpretation [6, 14, 31]. Prior work also indicates that data sharing within scientific teams is not always adequately supported by current technology [39]. Thus, it is clear that in order to further develop technology to support collaborative data sharing, we must first understand the challenges faced by academic scientists as they share data within their research groups.

In this paper, we explore the challenges associated with data sharing within academic scientific collaboration through interviews and observations of data sharing practices with 38 researchers from a variety of academic fields. Our preliminary study in this paper revealed that academic researchers tend to use off-the-shelf products (e.g., Google Drive and Dropbox) due to the increasing level of expertise needed to utilize sophisticated file-sharing systems, such as FTP sites. However, little research has focused on the challenges surrounding scientific data sharing within these types of collaborations when using off-the shelf systems, making this paper one of the first to tackle modern data sharing with this domain.

We identified key challenges in data sharing practices with regards to access control, and utilized the key insights from our preliminary study to design and implement *DriveGroups*. DriveGroups is an open-source system (shown in Figure 1) that facilitates data sharing between collaborators by allowing users to manage file access from two separate perspectives: the traditional file perspective and the role-based group perspective. These dual perspectives simplify the data sharing process by facilitating role-based, fine grained control. We further evaluated DriveGroups by performing a laboratory study with 18 researchers to directly compare DriveGroups and Google Drive with regards to usability, access control, and transparency. From these studies, we derived implications for the design of future data sharing systems.

Thus, the main contributions of this paper include:

- new knowledge surrounding the challenges and practices of scientists when sharing data within collaborations,
- a system, DriveGroups, designed to facilitate data sharing, and
- implications for the design of future data sharing systems.

2 Related Work

Most research on scientific data sharing has focused on sharing with non-collaborators [16, 20, 21, 24, 38, 46, 49, 50, 52], and the public [12, 18, 22, 23, 35]. Some additional research has focused on data sharing strictly from a security, rather than user-centered, standpoint [11, 25, 47, 48] or how data is structured, published, and discussed rather than shared [32]. As the focus of this project is on sharing within scientific collaborations, we concentrate on internal data sharing in this context, including challenges and existing tools.

2.1 Sharing Data within Scientific Collaborations

A large number of studies investigating scientific data sharing focus on public dissemination of data [23] or distribution to non-collaborators [21, 38, 52]. One study that focuses on data sharing within collaboration is the investigation conducted by Borgman et al. on the data sharing dynamics between scientists and technology experts affiliated with the Center for Embedded Network Sensing (CENS). Borgman et al. found that the data needs and applications of the two parties at CENS are complementary, competing, and interdependent [8]. Both teams required knowledge of the other's domain to implement accurate scientific standards and uncover patterns in the data. However, their differing research questions and methods resulted in a limited ability to recombine each groups' data for later use.

In the realm of planetary science, Vertesi and Dourish observed collaboration within two NASA mission teams [51]. Due to variations in attitudes regarding data collection methodologies, they found that attitudes towards data sharing varied between the two teams; one team viewed data as a group resource and freely shared the data with internal collaborators, while the other team was more cautious. Vertesi and Dourish concluded that the organizational culture dominating data collection is the key difference affecting data sharing.

Another example is the case study conducted by Buckeridge et al. on a community/university collaborative research project in Toronto, Canada. Buckeridge et al. developed a geographic information system for ready access to routinely collected health data, and studied the collaboration process that involved the use of such a system [10]. Among other findings, they mentioned that some data holders were reluctant to share their data, out of concern for privacy and data ownership once the data was released. To avoid adding to such concerns, data sharing tools should promote transparency and access control, so that users are clear about and have control over who has access to their data. With that said, the study focused on the collaboration between university researchers and community collaborators, and the developed system specialized in sharing of community health information. It is therefore different from our focus on collaboration within academia and tools researchers across disciplines use to share data. Similarly, work by Langmead and Nellore described the cloud computing model in collaboration, focusing solely on genomic data sharing and analysis [33]. Their work can hardly be generalized to inform how researchers who do not work with genomic data share data in practice.

The collaboration of data scientists and other science researchers has become an essential factor in investigating complex scientific problems. However, these partnerships often face challenges that call for a more user-focused design [36, 43]. Mao et al. analyzed the lack of common ground found between heterogeneous teams of interdisciplinary research teams. They found that a key challenge lies in the disagreement between what, how much, and with whom information should be shared [36]. Edwards et al. pointed out that metadata can lead to frictions for researchers with different backgrounds, thus impeding the collaboration process [19]. Jin and Ahn suggest that existing systems experience these breakdowns due to ambiguous data management controls, and

the resource owner should have complete authentication of all collaborators [27]. As such, access controls should be comprehensible for one person.

Prior work investigating user interfaces for data security software suggests existing systems are too complicated to be used effectively and often lead users to make mistakes when controlling access [28, 37, 43, 53, 56]. Smetters and Good [43] found that users with sharing rights rarely adjust access permissions on files and prefer to inherit the default permissions, suggesting confusion with existing systems. Additionally, collaboration in scientific domains is highly dynamic, and strict systems make editing permissions complicated [27]. Strict access control regulations by institutions and by the systems themselves lead data owners to desire more fine-grained settings to keep their data secure. A dominant issue is the lack of transparency in existing access control systems. Although users frequently interact with these systems to create or modify permissions, they largely reported difficulty in setting permissions, a preference for interface feedback, and desire for their documents to reflect the set permissions [53]. Johnson et al. [28] called for a system that prioritizes transparency in responsible parties for changes to both the data and the access permissions, a guarantee of security, and an easy-to-use interface.

These prior works suggest that future systems must promote transparency to allow users to see how their permissions are affecting their data. Collaborators require meticulous access control software to control the utilization, representation, and management of their data in order to work successfully. However, there are two deficiencies in this body of prior work. First, most of these studies are not current or recent, and therefore do not necessarily represent the status quo of data sharing. This is especially true considering the emergence and development of new data sharing tools. For example, most works have not discussed online sharing drives such as Google Drive, which, as Section 3 will show, were identified by our preliminary study as the most commonly used data sharing tool in academic collaboration nowadays. Second, prior works also lack an understanding of how attitudes towards data sharing affect the development of software. The study by Buckeridge et al. [10] might be an exception, but their investigation focuses solely on sharing community health information, without offering broader insights into data sharing in general. With these deficiencies in mind, we hope to provide a clear and relatively up-to-date account of data sharing challenges researchers encounter, as well as to bridge the gap between understanding attitudes regarding access control and deriving implications for the design of new systems to support data sharing.

2.2 Existing Tools for Data Sharing in the Sciences

In 2015, Morrison-Smith et al. [39] indicated that life scientists primarily utilize FTP/SFTP systems and physically mailing hard disks when they share data within collaborations due to the size of their files. However, prior work investigating access control within FTP/SFTP systems has been limited to the exploration of security and preventing unwanted access rather than usability [54, 55]. In the nearly ten years since the publication of Morrison-Smith et al.'s paper [39], off-the-shelf technologies such as Google Drive and DropBox have significantly increased their file-sharing limits. Furthermore, anecdotal evidence suggests that these are now popular methods for sharing data internally in the sciences.

Dropbox features a role-based sharing system with Dropbox Groups. This system allows an administrator to create, manage, and share folders with approved users; however, individual files cannot be shared with a group. This has the potential to lead to wrongful permissions sharing, where users are mistakenly given access to sensitive data. The system also does not have a method for viewing all folders shared with a particular group. This may lead permissions to remain with groups of users after their collaboration on a project has finished [17].

Google also seems to offer a potential solution with its Google Groups software; however, sharing files through this architecture is cumbersome. In order to share a file using Groups in Google Drive, the user must open Google Groups – a distinct software from Google Drive – and create a separate account, then create a group and manage accounts, and finally navigate back to Google Drive to share the desired content. The system lacks transparency, as users cannot view nor manage group members within Google Drive itself. In fact, the sole benefit of using Google Groups to share files is the facilitation of adding and removing emails. When managing file access, groups are presented as email addresses ("group_name@googlegroups.com") and members of each group are not indicated, which adds to the difficulty of confirming who has permissions. Therefore, the use of Google Groups could create barriers to transparency of the access different collaborators have, with transparency being a significant aspect of users' need for data sharing tools [29, 45].

Borgman et al. [7] studied the suitability of software-as-a-service technology (i.e., applications like Google Drive delivered over the internet, typically licensed as a subscription) for collaborative data sharing in small and medium-sized laboratories and concluded that the basic features of software-as-a-service tech are beneficial in terms of reducing costs, labor, and the need for tech support. Their study also identified several systems that are currently being used for data storage and sharing, including Dropbox. However, we suspect that these solutions may not be well-suited for managing the environmental information necessary for sharing scientific data. Thus, given the changing landscape of data sharing methods, there is a need to investigate data sharing in the sciences from a user-centered approach.

3 Preliminary Study: Exploring Access Control Challenges Within Collaborations

The goal of this study was to clarify the issues identified in prior work associated with access control when sharing data with collaborators in the context of scientific research. We conducted semi-structured interviews and observed scientists from a broad range of disciplines sharing data to investigate the mechanisms that result in a lack of transferable access settings, poor system visibility, and inconsistent ownership over files in institution-agnostic data sharing systems (e.g., Google Drive). By doing so we aimed to infer how scientists share data and understand their needs in order to develop our system. The primary purpose of the preliminary study was to understand the broader challenges and needs in group-based data sharing. As outlined in Section 4, we later developed a system named DriveGroups with awareness of the challenges researchers encounter. The results of this preliminary study thus informed the design of DriveGroups.

3.1 Method

Four investigators conducted semi-structured interviews [34] with scientists either pursuing or having completed a PhD in relevant fields and observed their software usage to examine the challenges faced when researchers share data within collaborations. While sharing files enables collaboration, the type and goal of scientific research dictates the nature of data sharing. Thus, we interviewed academic researchers from a variety of fields and asked them to contextualize data sharing with regards to their specific research goals. The participants in this study were 38 researchers aged 23 to 64 (M = 37.95, S.D. = 8.83, 19 female) from 14 universities, colleges, and research institutions located mainly in the United States, with some additional participants located in Canada, Australia, and China. Participants were randomly assigned a three digit identifier (e.g., P105), which was used to anonymize the data. We used data saturation [4] to establish our sample size (i.e., data collection was terminated once further sessions resulted in minimal new information). We recruited two participants who publicly made "Ask Me Anything" posts on science discussions

on Reddit [42]. The remaining participants were recruited via interdepartmental email or word-of-mouth. Each participant had or was pursuing a PhD in a scientific discipline (Table 2 in Appendix A) and was part of at least one research project involving data sharing across collaborators.

Interviews with 12 of the participants occurred in their primary workspace, one was conducted in a nearby empty office, and 25 occurred via Skype, Zoom, or phone. Interviews were between 22 and 63 minutes long (M=41.67 minutes, SD=11.23 minutes), and were recorded in audio format then later transcribed. To establish context, we first asked participants to describe their research, including the potential impact of their work, a description of their collaborations (e.g., the roles that they and their collaborators play), and the typical demographics of their collaborators (e.g., number of collaborators and their backgrounds). We also asked participants to discuss the software they use throughout their research process, including data sharing software (e.g., what tools they use, benefits of specific tools, problems using the tool). At the end of the interview, we asked participants to describe a hypothetical future technology that could be used to facilitate collaboration. This question was designed to prompt the participant to articulate their needs, rather than produce actual ideas for future technology. Our full list of interview questions is available in Appendix A. The protocol was approved by Institutional Review Boards at our three locations (IRB 201602178, IRB 20210505006, and IRB S22-002). Participation in interviews was voluntary and participants were compensated with $10 USD.

3.1.1 Data Analysis. We performed a bottom-up analysis of participants' responses by constructing an affinity diagram [3, 44] to identify prevailing themes in their research goals and work practices. This approach follows the qualitative analysis via coding as outlined by Auerbach and Silverstein [2]. However, instead of independently organizing data followed by calculating inter-rater reliability (IRR) as in qualitative coding, the four researchers analyzing the data came to a consensus on all responses. This approach is appropriate for semi-structured interviews, as qualitative coding results in the possibility of applying the same code to different sections of the interview [30]. We then examined themes from prior work [40, 41], which enhanced our interpretation of the interview data and allowed us to draw comparisons between our findings and prior knowledge, highlighting new discoveries.

3.2 Results

Although our participants mentioned utilizing a variety of systems for data sharing, for the purposes of this paper, we will focus our attention to cloud drives, as this was by far the most commonly utilized system for scientific data sharing with 84.2% of participants using Google Drive. With regards to cloud-based shared drives, we identified four main themes: 1) size and technical expertise drive platform choice; 2) inconsistent ownership over files; 3) lack of visible file sharing settings; and 4) the need for institution agnostic technology.

3.2.1 Size and Technical Expertise Drive Platform Choice. We found that overall, size and technical expertise were the primary drivers behind participant selections in data sharing platforms; users preferred to use systems that either were uniquely appropriate for the size and type of data being shared, or that they and their collaborators were familiar with. Sixteen out of all 38 participants stated that file sizes and types, as well as their field of work, played a role in their data sharing habits and choices of tools. For instance, P447 mentioned that although they sometimes used GitHub for sharing data, it was "not great" for large files. They therefore used other tools when their research involved such files. Seventeen participants pointed to familiarity and technical expertise as drivers behind their data sharing choices. These included not only their own comfort levels with practices and platforms, but also their perceptions of others' expertise and preferences. The majority of participants reported collaborating with other academic researchers from other fields.

P105 in particular expressed that their collaborators were often less technically adept due to having biological rather than computational backgrounds, and as a result P105 often preferred emails and shared drives over FTP servers. They remarked that some collaborators "just don't know how to get on the server." Our participants reported relying on three main types of tools: 1) cloud drives (e.g., Google Drive, Dropbox, and One Drive), 2) systems designed primarily for messaging (e.g,. email, Slack, Teams, Discord), and 3) file servers with either FTP or virtual machines set up to share data across institutions. All but one participants had used at least one of these tools to share data.

3.2.2 Inconsistent Ownership Over Files. We also saw that researchers must perform balancing acts between controlling access and facilitating collaboration, as well as between convenient data sharing practices and protecting data from non-collaborators. Concerns about control over data and documentation had a profound effect on the research workflow and tools for sharing data—both in terms of comfort working with collaborators and restrictions on the influence of a collaborator on a project. Thirty-two out of all 38 participants noted that they would limit access to data even within collaborations, sharing only parts of their data with the appropriate collaborators. We saw this trend across disciplines. The primary reasons for wishing to limit sharing to subsets of collaborators included: 1) preventing students from accidentally affecting more files than they're supposed to; 2) reducing the number of automatic update notifications for high level collaborators to reduce annoyances; 3) limiting student access to sensitive de-anonymized data; and 4) limiting the number of people who can modify files at one time, such as manuscripts. When sharing data about human subjects, data would first have to be anonymized to prevent sharing of personal patient information. To control access, participants mostly used built-in features of data sharing tools such as Google Drive, Dropbox, and GitHub, while seven participants also mentioned using alternatives that they felt were more secure. For instance, P678 said that their students could access data only from lab machines and not from their own computers, citing security as the main rationale. Additionally, six participants noted that they had relied on experts to set their systems up, especially when it came to more complex data sharing tools such as file servers. In an effort to alleviate problems controlling access to data, especially with multiple collaborators, 13 participants organized data into root folders representing individual projects. Each root folder then had sub-folders which a participant could modify access of on an individual-by-individual basis. However, within such a system, problems originated from various access level needs for particular files, setting up and enforcing a file management structure, and difficulty traversing file structures especially in Google Drive. It is also worth noting that reflecting upon their data sharing processes during the interview, four participants realized that they had less control over their data than they had previously thought, pointing to a lack of transparency about who could access their data.

3.2.3 Lack of Visible Sharing Settings. Systems often lacked visible default data sharing settings that facilitated setting and transferring permissions. Furthermore, six participants expressed concern that they would accidentally share data with their collaborators when using shared drives because of the passive nature of the sharing (i.e., automatic sharing). This concern about accidental sharing is problematic since as 32 participants mentioned, users *"don't necessarily want everyone to have access to all the data"* (P254), even though they were sharing with collaborators as opposed to sharing publicly. The problem could be remedied with more visible sharing settings. In addition to difficulty granting access to collaborators, 15 participants pointed to challenges with visibility of access in existing systems, such as forgetting to grant/remove access to a file, having access expire, lack of clarity on the status of access, and unintentional access level changes such as when a student graduates and is no longer granted access to university resources. In addition, participants mentioned difficulty providing access, such as not knowing which email(s) belonged to which

collaborator, and the fear of accidentally sharing the wrong files with the wrong individuals who simply have the same name or similar email address to a collaborator.

3.2.4 Institution Agnostic Technology. Twelve participants also noted that it was common for a collaborator's institution to deny researchers from other institutions access to files. As a result, whenever possible, participants chose to use data sharing technologies that were institution agnostic and thus, accessible to non-co-located collaborators (e.g., email, Google Drive, Dropbox, or mailing physical drives). Other times, researchers were forced to find an alternative method, such as mailing physical hard drives with the data. Thus, it is clear that any system that we expect to have widespread adoption amongst scientists must be institution-agnostic. This is important given that the majority of our participants routinely worked with researchers at other institutions, sometimes (like in the case of P108) to ensure that a project has a CO-PI with specific expertise, and other times as a result of moving to another institution, such as the case of P927 who maintained collaborations with researchers at the institution they received their PhD from after securing a faculty position at a new institution.

3.3 Discussion

Our preliminary findings validate prior work by Koesten et al. exploring internal data sharing in the context of a broad range of domains (e.g., public administration, education, and finance) which showed that collaborators struggled with the need for version control and controlled access [32]. Additionally, work from Jin and Ahn [27] proposes role-based access control as the solution to delegate permissions in these collaborative settings to alleviate sharing mistakes and provide access transparency. From our findings, we identified three key instances when researchers nowadays encounter challenges sharing data: 1) sharing across institutions, 2) sharing across domain or research field, and 3) sharing across level of expertise. We derived the following implications for the design of software for sharing medium sized files within research collaborations within these three themes. These implications arose from our participants' diverse collaborative processes.

3.3.1 Sharing Across Institutions. Our preliminary research and prior work [39] indicate that it is common for academic research collaborations to span institutions in order to provide access to specific expertise, such as in collaborations containing both computer scientists and life scientists, and to preserve preexisting working relationships when a researcher changes institution. For the purposes of our design recommendations, this is what we mean by sharing data across institutions, and we focus on two main themes: multi-institutional sharing and setting project defaults.

Multi-Institutional Sharing: Any system designed for collaborative research should support sharing across multiple institutions. This need arose from our participants' collaborative processes, as they often chose to work with collaborators at other institutions, a finding that validates Morrison-Smith et al.'s 2015 paper [39]. Supporting multi-institutional sharing ensures broad accessibility and facilitates collaboration across various organizations, breaking down barriers that might be caused by licensing restrictions.

Setting Workspace or Project Wide Defaults: Systems for collaborative data sharing should incorporate methods for setting defaults at the workspace or project level. This need arose from the collaborative research process, where participants discussed setting up file structures on a semester-by-semester basis for multiple teams. Typically, the composition of these groups remains consistent (e.g., PIs, CO-PIs at other institutions, and students), although the individual members may vary. Implementing standardized defaults streamlines the setup process, ensuring consistency and efficiency in system usage across different teams or projects. This approach also helps maintain standard practices within and across institutions.

3.3.2 Sharing Across Domains. In addition to collaborating with researchers at other institutions, prior work [39] as well as our preliminary results indicate that it is common for academic researchers to participate in multidisciplinary collaborations. Since some fields, like computer science, tend to be more tech-savvy than others, these collaborations commonly have a mixture of both research and technological expertise. In this context, we provide two recommendations for the design of systems that facilitate data sharing in collaborations that span research domains: improve transparency and discoverability, as well as focus on usability.

Improving Transparency and Discoverability: Enhancing the overall transparency and discoverability of the system is vital. Seven of our participants expressed uncertainty whether some feature they needed was already incorporated into a tool they used, or were simply unaware of the existence of such a feature because they came from fields where the focus of instruction was not computational. Users should find it easy to navigate and locate the information or tools they need. This implies need for an intuitive interface, efficient search mechanisms, and well-organized content.

Focus on Usability: Systems should be designed with a strong emphasis on usability, ensuring they are user-friendly, intuitive, and cater to users with different levels of technical proficiency. This need arose from the collaborative research process, where varying expertise levels are primarily due to the interdisciplinary nature of large research projects where it is common for a computer scientist to work with an expert in another domain. A highly usable system encourages wider adoption and more effective use, particularly as the size of data being shared increases.

3.3.3 Sharing across skill levels. Our preliminary research indicated that academic research often incorporates a pedagogical aspect where senior PIs often collaborate with junior researchers, such as undergraduate and graduate students. Thus, research collaborations often consist of researchers who have varying expertise levels due to differing levels of education and experience. For these collaborations, we suggest three design implications: use or augment off the shelf technology, provide fine-grained controls, and ensure that sharing indications are visible and/or transparent.

Use or Augment Off-The-Shelf Technology: To address the issue of varying expertise levels among users, systems should leverage off-the-shelf technology that researchers are already familiar with. Leveraging familiar technology minimizes the learning curve and allows for smoother integration into existing workflows. Additionally, integrating with established technologies increases the likelihood of ensuring compatibility with other tools and platforms that researchers frequently use.

Fine-Grained Controls: Such systems also need to provide users with fine-grained control over various aspects, including detailed permission settings, data management, and customization options. This need arises from the hierarchical structure of collaborative research teams, where PIs and CO-PIs require full access, while students often have limited access. In some cases, PIs have full access, CO-PIs have limited access (only to papers and finished data), and students have access only to raw data and finished data. Role-based access control, as noted by both prior work [27] and participants, allows users to tailor the system to their specific needs, ensuring a secure and personalized user experience.

Visible/Transparent Sharing Indications: The user interface of a collaborative data sharing system must include clear indicators of how and when data is being shared. This can be achieved through visual cues, notifications, or logs that inform users about shared data, access levels, and any modifications made. This need arises from the pedagogical nature of academic research, where participants often need to remove students from projects once they have graduated or left the project. Participants commonly noticed that graduated students remained on projects they should no longer have access to. Such transparency is essential for maintaining trust and ensuring effective collaboration.

3.4 Limitations and Future Work

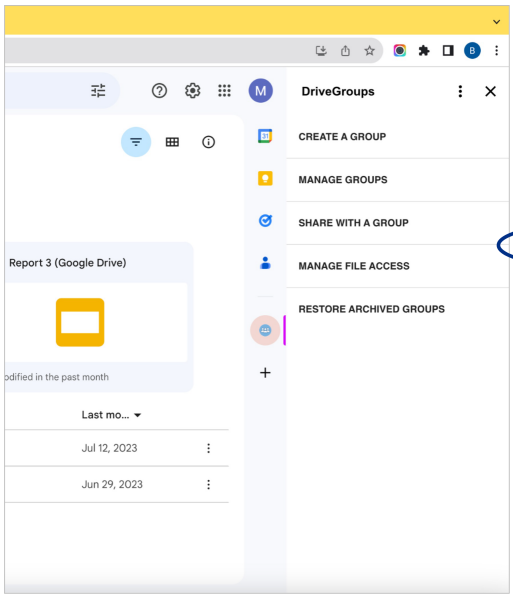
While this study has provided valuable insights into the challenges and needs of our participants in managing data, our sample predominantly consisted of individuals from academic institutions, which may not fully represent the broader spectrum of data management practices used in various industries. Many organizations outside of academia utilize internal or enterprise tools, and their experiences may differ significantly from those of our participants. Additionally, while our study allowed us to gather in-depth information about the data management practices of participants who did not face significant issues related to data size, it might not fully capture the experiences of individuals and organizations dealing with larger datasets. Although a few participants reported working with data large enough to require the use of FTP servers and physically mail hard drives, the majority of our participants reported using small enough files that they were manageable with existing cloud-based systems. Consequently, the challenges associated with handling extremely large datasets were not a central focus of this research. Future work will consider addressing the needs and concerns of participants from a more diverse range of sectors, including those dealing with substantial data volumes, as this remains a pressing issue in data management.

4 DriveGroups: Facilitating Access Control in Google Drive

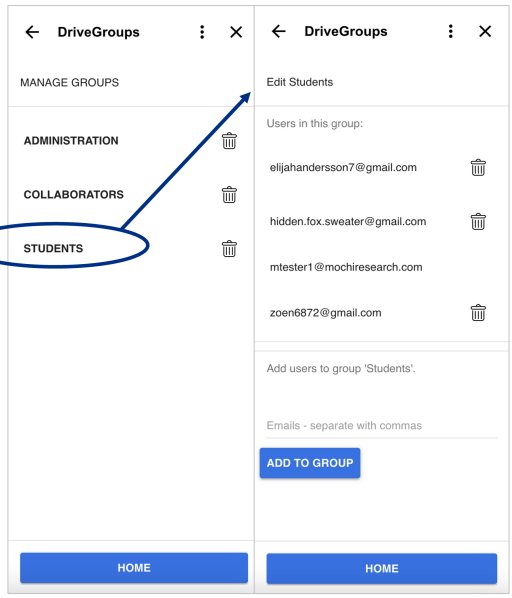
Based on the feedback from our preliminary study, we developed a system called *DriveGroups*. This system augments Google Drive's features, as Google Drive was the most popular system for internal data sharing used by our participants (used by 84.2% of our participants). Google Drive is also easily extendable using AppsScript, a derivative of JavaScript with built-in Google API support, and integrates well with the SQL database provided by Google Cloud. DriveGroups aims to improve data sharing by addressing unmet user needs such as increasing sharing settings visibility, supporting the transfer and reuse of access control settings, and standardizing ownership over files which persist in off-the-shelf systems. The primary mechanism is the allowance of managing file access from two separate perspectives: 1) the file perspective and 2) the role-based group perspective (see "Manage File Access" in Figure 2). The file perspective is similar to file access management in Google Drive; after selecting a file, users can view and edit the list of collaborators with access. Meanwhile, from the group perspective, users can manage a selected group's access to all relevant files. This perspective is more efficient when, for instance, the user needs information on which files a group has access to. DriveGroups accomplishes this by constructing a query to the MySQL database running on a Google Cloud Platform virtual machine dependent on the current sharing perspective, the individual(s) being modified, and file information. A tutorial video showcasing DriveGroups can be accessed in the supplementary material.

4.1 Sharing Settings Visibility

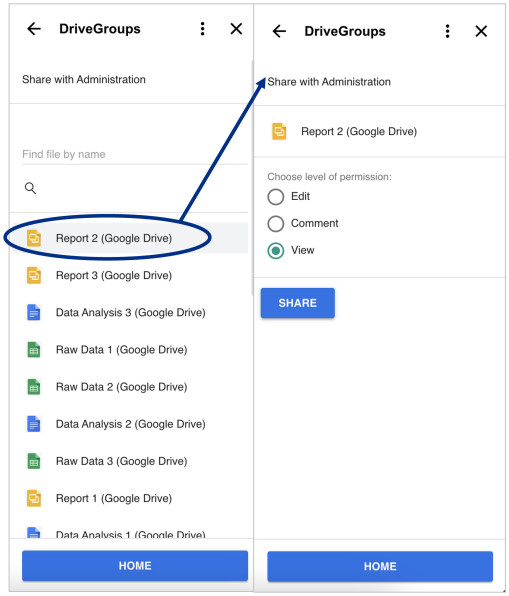
Our preliminary work identified that a lack of visible permissions leads to concerns regarding accidental sharing. In Google Drive, settings concerning sharing are hidden from view and must be accessed through a series of menus. We redesigned the user interface to highlight information depending on the sharing perspective chosen. For a selected file in the file perspective, the user interface illustrates which group has access to a document, the level of access given to each group, and allows the user to modify a file's permissions for each group. For a selected group in the group perspective, the user interface displays all files shared to that group, the permissions the group has for each file, and lets the user change the groups file permissions by file (see Figure 2). DriveGroups consolidates this information dynamically via batch requests as the user browses and interacts with the system. By presenting data in this way, we hope to avoid hidden and unclear representations of relevant information to their tasks in every window.



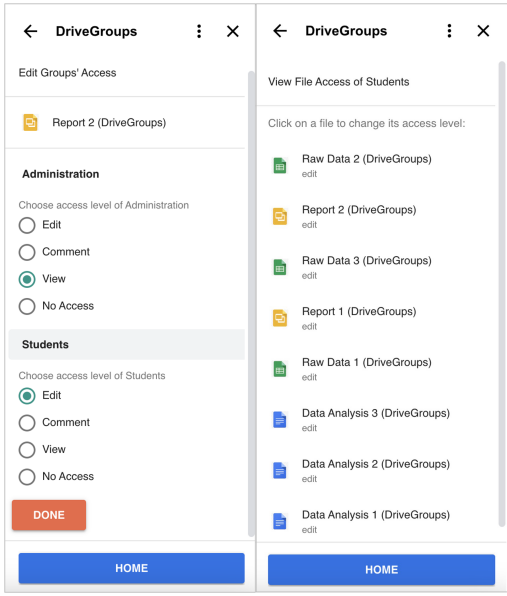
Main Menu



Manage Groups



Share with a Group



Manage File Access

Fig. 2. Within DriveGroups users can easily add collaborators and share specific files to each group. With one click on the Main Menu, users can manage group members, share files with a group, and control file access for a selected group.

4.2 Supporting the Transfer and Reuse of Access Control Settings

Our preliminary study revealed that Google Drive's existing architecture falls short in several ways in regards to access control. The most prevalent example of such is the lack of ease users have with

sharing files with large groups of individuals with varying access needs. In Google Drive, users have to modify file access for each individual which is tedious. While Google Groups exists, it is not user-friendly nor sufficient for these issues. The drawbacks discussed in Section 2.2 possibly explain why no participant in our preliminary research indicated knowledge of or experiences with Google Groups.

In contrast, as an add-on to Google Drive, DriveGroups opens within the Google Drive interface and allows the user to create, view, modify, and share with groups all in one place (See "Main Menu" in Figure 2). Furthermore, since it interfaces as a side-bar add-on instead of an external service, DriveGroups integrates well with Google Drive's ecosystem. Users are able to utilize access control features from Google Drive and DriveGroups interchangeably in their workflow. As discussed in section 4.1 and section 4.3, DriveGroups is able to provide these access control features without sacrificing information transparency in file visibility and ownership.

4.3 Standardizing Ownership over Files

We identified in our previous study that with file-sharing software such as Google Drive, users frequently deal with ownership issues such as a lack of sharing awareness. To address the various privacy needs for sensitive and/or confidential information, the design of DriveGroups allows users to easily view and track shared files and levels of access at any given time as long as the user has a method of accessing Google Drive. We accomplished this by segmenting relevant data in a MySQL database. We split our data into two segments: group demographics and file data. Group demographics, while stored in a separate table from file information, contains the unique IDs of each file (which Google Drive generates upon creation) that a group has access to. Thus, modifying a group does not modify a file's properties and vise versa. However, group demographics do store their the respective access level with their accessible files. This allows multiple groups to access the same file but from different levels of access.

As the owner of their files, users can use DriveGroups to create groups and grant or rescind access to their information at any time and in mass. When a user updates access levels from the group perspective (see "Share with a Group" in Figure 2), DriveGroups iterates through each email associated with the affected group and updates their access in Google Drive. Thus, not only are the changes visible in DriveGroups but also in Google Drive. When the user updates access levels from the file perspective (see "Manage File Access" in Figure 2), the group ID (created by DriveGroups at the time of group creation) stored in that particular file's entry in the database is used to reference the affected group. DriveGroups then repeats the algorithm used in the group perspective to modify access levels.

4.4 DriveGroup's Contributions

The development of this system is motivated by a wide range of challenges, including data security, access control, transparency, visibility, ease of use, version control, and accidental sharing. Our system specifically offers value in terms of access control and managing data sharing. In Table 1 below, we delineate these different concerns and clarify what we mean by security and data protection. By addressing these concerns through specific system features, we aim to provide a robust solution that enhances data security, access control, and overall efficiency in collaborative research environments.

Table 1. DriveGroup's contributions toward security and data protection.

Category	Definition	System Contribution
Data Security and Protection	Data security refers to measures designed to protect data from unauthorized access and breaches, ensuring data integrity and confidentiality.	Our system implements role-based access controls to ensure that only authorized users can access sensitive information.
Access Control	Access control involves regulating who can view or use resources in a computing environment.	The system allows detailed permission settings, enabling users to define who can access specific data and to what extent. This is crucial for maintaining the hierarchical structure of research teams, where access needs vary among PIs, CO-PIs, and students.
Transparency and Visibility	Transparency involves making data sharing activities visible to all stakeholders, ensuring everyone is informed about data access and modifications.	DriveGroups provides clear indicators of shared data, access levels, and changes made. This transparency helps in maintaining trust and effective collaboration, particularly in ensuring that only current team members have access to ongoing projects.
Ease of Use	Ease of use refers to how user-friendly and intuitive the system is for users with varying technical proficiency.	By leveraging off-the-shelf technology and familiar interfaces, the system minimizes the learning curve and facilitates smoother integration into existing workflows. This is especially important in academic settings with mixed levels of technological expertise.
Version Control	Version control involves managing changes to documents and data over time, ensuring that previous versions can be retrieved if necessary.	The system includes version control features that track changes, allowing users to revert to previous versions if needed, thereby preventing accidental loss of data.
Accidental Sharing	Accidental sharing refers to unintended distribution of data to unauthorized users.	By incorporating detailed permission settings and transparency features, the system helps prevent accidental sharing, ensuring that data is only accessible to intended recipients.

5 Evaluating DriveGroups

5.1 Method

To evaluate DriveGroups, three researchers conducted a series of usability and basic performance evaluations with 18 research assistants aged 18 to 23 (M = 20.11, S.D. = 1.13, 9 female, 7 male, 2 gender-expansive) at a local institution in the United States. There was no overlap between the participants in the first study and the second study. We used data saturation [4] to establish our sample size (i.e., data collection was terminated once further sessions resulted in minimal new information). Participants were recruited via interdepartmental email and word-of-mouth. Participants were randomly assigned a three digit identifier (e.g., P105), which was used to anonymize the data.

Participants were given a sample project with nine files of sample data that they were expected to share with several fictional colleagues located at a variety of institutions and roles. These roles included: student researchers from the participant's lab, collaborators from other institutions, and funding administrators. These roles were chosen to reflect the types of collaborators our participants from the preliminary study reported sharing data with. After watching a six-minute tutorial explaining how DriveGroups works, participants were encouraged to "think aloud" while using either DriveGroups or Google Drive to complete a series of data sharing tasks with the fictional colleagues. These tasks, as shown below, are designed to test if DriveGroups satisfies two major desires of researchers when data sharing: the ability to accurately share files to different groups of people and to easily change permissions, individually or collectively, for different files.

(1) Share your documents with your students, collaborators, and funding administrators at the appropriate access levels.
(2) Your collaborators work extensively with you one month and ask to contribute to report 3. Allow them to edit this document.
(3) Student A just graduated, remove his access to your files.
(4) You hire a new student. Their email is [redacted]. Give them access to files they need.
(5) You just discovered report 3 is from the wrong month, unshare it with your collaborators and administration.
(6) You've finished your research, remove access for all of your collaborators.

The findings from our first study showed that researchers have major security concerns about sharing confidential data. A mistaken grant of a lower level of data access could lead to a leak of participants' personal information. Moreover, an unintended grant of a higher level of access could hinder collaborators' ability to capture the big picture of the data. Therefore, tasks (1), (2), and (4) were designed to test whether DriveGroups decreases the chance of making such mistakes. Additionally, manually monitoring inactive emails in labs can be time-consuming and may lead to mistakenly removing current members in large-scale projects. Tasks (3), (5), and (6) focus on testing DriveGroups' ability to reduce the laborious work required to modify file access for individuals and groups.

Participants were not presented with a tutorial for Google Drive, as all participants reported having at least one year of prior experience using that system. The order in which participants used each system was counterbalanced to mitigate ordering effects. Usability was assessed using Brooke's System Usability Scale (SUS) [9], followed by a survey (listed in Appendix B) to gauge users' perceptions of each system's ability to control access and ensure transparency in sharing settings. After these quantitative assessments, participants were asked a series of semi-structured interview questions (listed in Appendix B) to delve deeper into the specific aspects of the tools that led to positive or negative outcomes. This sequence—starting with the SUS and survey, then moving to qualitative interviews—allowed us to understand not only how participants rated the systems but also why they rated them in that way. This approach provided a comprehensive view

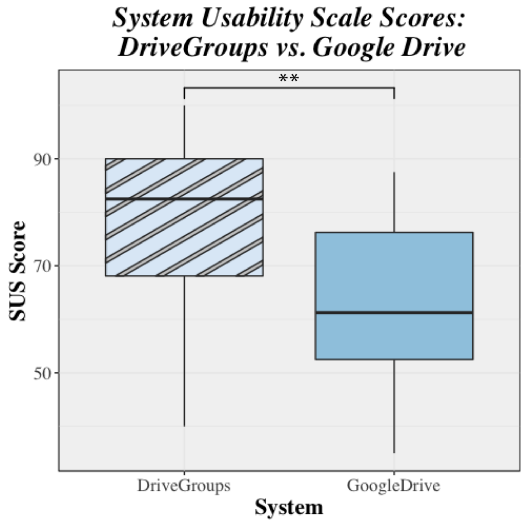


Fig. 3. System Usability Scale scores for DriveGroups vs Google Drive. ** Indicates significance at p < 0.001.

of user experiences, linking quantitative scores to qualitative insights. Although it is more common to begin with broad qualitative questions to avoid anchoring participants with specific interests and ideas, we chose this methodology to first gather structured, comparable data through the survey. The subsequent interviews then explored the reasons behind these ratings, ensuring a nuanced understanding of user perceptions and experiences. This order might introduce limitations, such as anchoring participants' thoughts with specific survey questions before the interviews, but it was essential for correlating the quantitative data with qualitative insights. Video recording captured both the participant's face and their screen as they completed the tasks. This study took between 53 and 90 minutes (M=65.28 minutes, SD=8.05 minutes). This protocol was approved by our Institutional Review Board (IRB S22-003). Participation in the evaluation was voluntary and participants received compensation of $25 USD.

5.1.1 Data Analysis. As with the preliminary study, the "think aloud" comments and responses to semi-structured interviews were analyzed by performing a bottom-up analysis of participants' responses by constructing an affinity diagram [3, 44] to identify prevailing themes explaining our quantitative results and identify what aspects of the system are promising for the final version, which will be released open-source. This approach follows the qualitative analysis via coding as outlined by Auerback and Silverstein [2]. However, instead of independently organizing data followed by calculating inter-rater reliability (IRR) as in qualitative coding, the five researchers analyzing the data came to a consensus on all responses. This approach is appropriate for semi-structured interviews, as qualitative coding results in the possibility of applying the same code to different sections of the interview [30].

5.2 Results

5.2.1 Usability. All 18 participants evaluated both DriveGroups and Google Drive using John Brooke's System Usability Scale [9]. DriveGroups (M = 78.25, SD = 16.86) outperformed Google Drive (M = 59.88, SD = 16.89) in terms of usability, t(38) = -3.44, p = 0.001. Results are shown in Figure 3. Participants offered positive feedback regarding DriveGroups' usability. Sixteen out of 18 participants described DriveGroups with expressions such as "easy," "friendly," "convenient," and "fun to use." Participants liked how DriveGroups helped to prevent and recover from errors. Five participants stated that with the help of groups, they did not have to share individually with

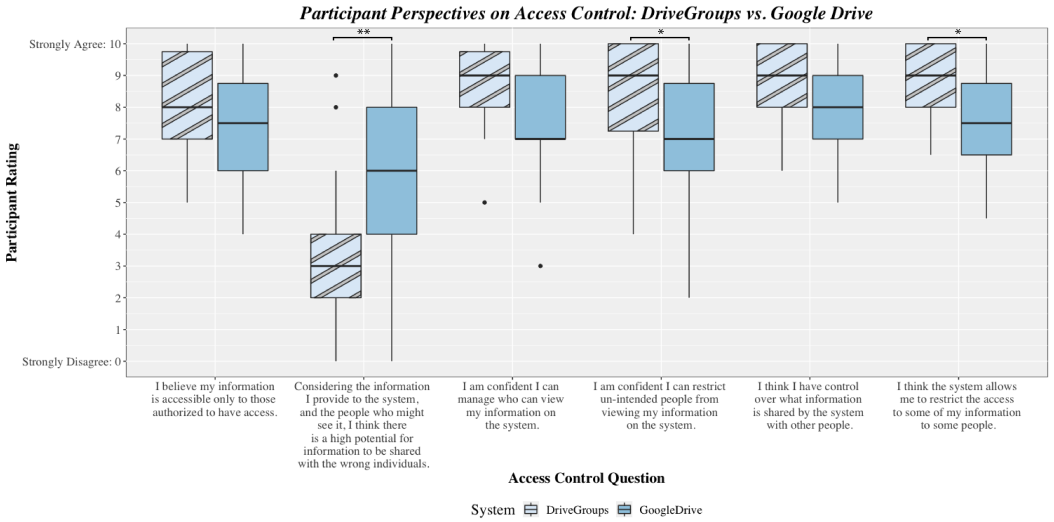


Fig. 4. Responses to access control questions for DriveGroups vs Google Drive. * Indicates significance at $p < 0.05$, ** indicates significance at $p < 0.01$.

collaborators, which simplified the sharing process and reduced mistakes that were more frequent in repetitive individual sharing. Seven participants spoke positively of DriveGroups' design in terms of error prevention. This included, but was not limited to, the ability to easily undo actions and recover archived groups.

Ten participants stated that they felt DriveGroups helped them complete tasks faster and with less effort compared to Google Drive. DriveGroups' group feature also resolved the issue of repetitiveness that 12 participants encountered in Google Drive when inputting individual email addresses for each file that needed to be shared. Six participants mentioned that with DriveGroups, they did not need to repeatedly check the level of access each collaborator should have, since the organization and names of groups already incorporated this information and made it easy to track who already had or still needed access.

Finally, participants noted that when they needed to work with a large number of files or collaborators, DriveGroups would work especially well compared to Google Drive. Eleven participants noted that sharing with Google Drive became extremely tedious when they had more files and collaborators to work with, and they expressed that DriveGroups would be more helpful in such cases.

5.2.2 Access Control. All participants rated six statements about access control (shown in Figure 4) on a visual analog scale [13] from 0 (Strongly Disagree) to 10 (Strongly Agree). Generally, DriveGroups met or outperformed Google Drive in terms of access control. Results are shown in Figure 4. Specifically, DriveGroups was ranked significantly better than Google Drive for the following three questions (note that question 1 is a flipped question):

(1) Considering the information I provide to the system and the people who might see it, I think there is a high potential for information to be shared with the wrong individual.
(2) I am confident I can restrict un-intended people from viewing my information on the system.
(3) I think the system allows me to restrict the access to some of my information to some people.

Participants found access control convenient and reliable in DriveGroups. Eleven participants described DriveGroups' access control features as "easy to use," "helpful," or "reassuring." Five participants noted that when managing access with DriveGroups, they felt "in control" and did not

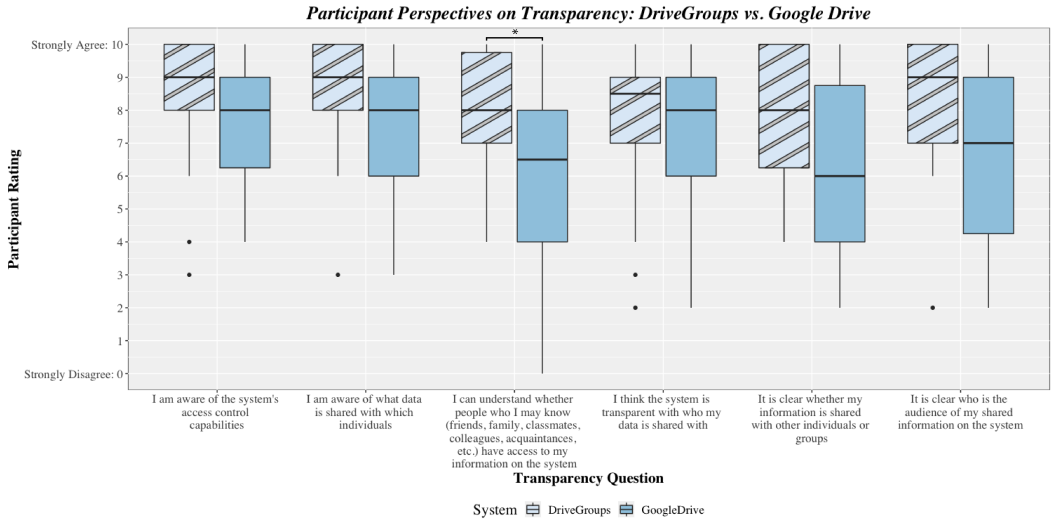


Fig. 5. Responses to transparency questions for DriveGroups vs Google Drive. * Indicates significance at $p < 0.05$.

worry about granting access to unintended individuals. One participant attributed their reassurance to the group feature, saying that groups "made sure the user knows who each file is going to (P614)." Another participant (P652) expressed similar sentiments and stated that groups alleviated the concern that they might forget individuals that they needed to share their files with.

Participants spoke positively of the features that enabled them to choose between managing access by file and by group. Four participants preferred to manage access control from the file perspective, mentioning it was easy to control all groups' access to a file in one place. Meanwhile, six participants preferred to manage access via the group perspective, stating that the feature, which they recognized as lacking in Google Drive, facilitated access management and control. In addition, four participants noted they liked that DriveGroups had both of these perspectives enabled, which offered two different ways to manage access. Participant P250 noted:

> "So the fact that you can check the access of files by group and by files for it helps you double check if you made any mistakes." (P250)

Participants put forward suggestions regarding DriveGroups' access control process. When a user wanted to change access of a group, DriveGroups set the default access level to change to "No Access," which aimed at minimizing the chance of undesired access granting. However, nine participants found this setting problematic, with four of them suggesting changing the default to the level of the access that the group originally had. Also, six participants were confused about whether removing a person or group would revoke their access to files. In practice, DriveGroups would revoke access an individual gained from a group when the individual was removed, and would revoke all recorded access a group had when the group was deleted. Participants stated that this was not made clear to them and suggested adding notifications or other types of clarification.

5.2.3 Transparency. Additionally, all participants rated six statements about transparency (shown in Figure 5) on a visual analog scale from 0 (Strongly Disagree) to 10 (Strongly Agree). Results are shown in Figure 5. Generally, DriveGroups performed equally as well as Google Drive in terms of transparency. However, on one question (*"I can understand whether people who I may know (friends, family, classmates, colleagues, acquaintances, etc.) have access to my information on the system"*) participants rated DriveGroups' transparency as higher than Google Drive's.

Participants found DriveGroups' transparency to be satisfactory. Seven participants stated they had little or no concern with DriveGroups' transparency, because they could easily check the level of access each group had to files. Four participants noted that the "Manage Access by File" and "Manage Access by Group" features contributed to DriveGroups' transparency by assuring the user that only selected groups had access to a certain file. Additionally, five participants stated that the ability to make and manage groups made DriveGroups more transparent. According to one participant, the list of people who had access could get "messy" in Google Drive. Comparatively, DriveGroups clearly listed group access whilst keeping track of group members.

Participants' opinions towards Google Drive's transparency, in comparison to DriveGroups, were mixed. Two participants explicitly complained about Google Drive's transparency, one of whom said that "it was murky who had access" (P603). In contrast, three participants believed Google Drive was transparent, noting that it was clear who had access to a file. Meanwhile, participants strongly stated that Google Drive made checking individual access levels tedious. Although three participants felt it was easy to check access, nine participants noted that access checking was tiresome because they needed to manually go through lists of files and emails to see if an individual had access. One participant found that the tedium made Google Drive less transparent; another mentioned that when checking access they were more likely to make mistakes in Google Drive, especially when the people who required access had similar names or emails.

5.2.4 Error Handling. A majority of participants stated DriveGroups provides various methods for handling errors which they felt gave them more control over their data than Google Drive does. In particular, nine participants stated the system's design and features prevented several mistakes from happening. These features included presentation of all group and file information when performing specific tasks, the separation of access control and sharing features into different windows, and the ability to undo most actions performed immediately after executing them.

Furthermore, 10 participants explained error recovery features like the "Restore Archived Groups" window is "convenient" and "good" for reverting the action of deleting a group in comparison to re-constructing the group from the "Create a Group" window.

6 Discussion

6.1 File-Sharing Improvements from DriveGroups

Overall, DriveGroups's design successfully provided better utility, access control, and sharing features when compared to Google Drive. This is due to our implementation of a system that facilitated interactions with groups instead of individuals. Our system has also addressed most of the issues raised from previous work and within our preliminary study regarding inconsistent file ownership, lack of visible sharing settings, and preference for institution agnostic systems. The reasons for such improvements is multi-faceted.

Off-the-Shelf Access Control: DriveGroups allows for various small to medium file formats, is transparent, and tracks file ownership to minimize confusion. In addition, DriveGroups is an extension of Google Drive, which many researchers have access to due to its relative institution agnosticism. DriveGroups also provides role-based access control that allows users to customize groups of contacts and sharing settings. These improvements thus address concerns from our preliminary study: preference for institutionally agnostic systems, lack of transparency over file ownership, and need for fine-grained access control.

6.2 Implications for Design

This study has important implications for the future development of DriveGroups as well as future data sharing applications. These implications are derived from participants' feedback on

design of DriveGroups regarding access management perspectives, transparency and sharing, and connections between permissions and groups.

Emphasize Accessibility for Protecting Data: We saw that researchers are not always informed about the methods they could use to control access to their data and do not necessarily understand the effect that different access control approaches would have on the availability of their data. It is therefore important that future technology prioritizes transparent mechanisms, and makes it easy for users to understand and employ its data control methods. It is also important to facilitate understanding of how to apply data protection techniques and how those techniques will affect the accessibility of data to collaborators and to the public. This will reduce non-computer scientists' reliance on the skills of computer experts to share their data.

Support Fine-Grain Levels of Control: Our participants' concerns about data control showed that researchers need technology to support fine-grain control over who can manipulate the data. In addition to being discoverable and documented, the data access controls also need to support nuanced management of access permissions. By helping researchers tailor permissions to their specific project and collaborators, new technology can facilitate scientists' feelings of complete control over their data and reduce fears that the integrity of their data will be damaged. We also know that scientists are more comfortable sharing data when they can specify conditions on access. These restrictions could require that collaborators only make their data available to the public under the same access conditions. New technology must include ways for scientists to communicate these stipulations for access. It is likely that by ensuring that their collaborators will respect their data sharing preferences, researchers would feel more agreeable about sharing data with their collaborators.

Enable Access Management from Multiple Perspectives: DriveGroups allows users to manage file access from two separate perspectives: 1) the traditional file perspective and 2) the group perspective. This is an example of how categorizing available tools and operations into multiple perspectives future systems can offer users flexibility for different specific tasks. Allowing choices in sharing perspectives can relieve the need for more complex file or group tools because more complex operations could be circumvented by viewing the task from the other perspective (e.g., in DriveGroups, mass-sharing a single file with a group of collaborators is much simpler from the group perspective than the file perspective). This type of categorization can also help organize options available to a user and make them more visible than simply displaying them all on one screen for a user to search through; thus addressing our participants' concern regarding lack of transparency in data sharing. By implementing options for multiple perspectives, systems can stand to gain greater usability and simplicity.

Ensure Transparency During the Sharing Process: Users need relevant information to be visible throughout the sharing process, not just before and after sharing files. Prior work [43] and our preliminary study revealed that when using data sharing applications, scientists need easier access to information such as who has permissions to some data. Our participants clearly indicated that when sharing files with a group they needed more information to be immediately available, such as the group's members or files the group already has access to. This finding emphasizes the need for transparency during sharing, which should not be confounded with verifying who has access after sharing is done.

Provide Information About Group File Access Relationship: Finally, future applications should take care with the relationship between groups and file access. DriveGroups' group sharing perspective was created to facilitate access control, and as a response to the need for consistent permissions and clear sharing settings as identified by both Johnson et al. [28] and our preliminary study. Nevertheless, while participants found the group creation and file access management processes to be intuitive, multiple participants asked whether adding or removing someone in a group would

correspondingly grant or revoke access. Similarly, after deleting a group, some participants were worried that previous members would still have permissions gained through the deleted group. These findings indicate that changes in access along with edits of groups could seem unintuitive to users. Possible solutions include providing information in the user interface that illustrates the relationship between groups and file access, but further research is needed to balance among easy access control, consistent permissions, and intuitive design.

6.3 Limitations and Future Work

Our work is limited by the biases present in our participant pools for both studies. Most participants were academics located in the United States, with a few from other countries. Our findings may not fully capture usability and performance issues that could arise in non-academic settings or with industry professionals. Future research will aim to include a more diverse range of participants, including those from industry and other non-academic fields in a wider range of locals, to better understand how DriveGroups performs across different contexts and user groups.

Additional future work includes iterating and refining DriveGroups. The participants in our usability study were primarily undergraduate researchers, lacking the experience of the participants from our preliminary study, who were performing a facsimile of collaborative tasks. This limits the results to the user interface's usability and perceived security and transparency, as participants were not engaged in actual collaborative work, not fully capturing the complexities and dynamics of real-world scientific research collaboration. In addition, more experienced researchers might focus on different aspects of the tool compared to less experienced researchers. To address this limitation, we are conducting an in-situ remote diary study with experienced academic researchers. Findings from our final in-situ evaluations will inform the design of the final version of DriveGroups, which will be shared publicly as an open-source Google Drive add-on.

Furthermore, because there was no overlap in participants between our two studies, the demographics for both studies were slightly different. The first study could be conducted remotely, allowing us to recruit academics with high levels of research experience. However, the in-person evaluation of DriveGroups was limited to locally available researchers. This presents possible limitations in terms of experience bias, learning curve, and evaluation metrics. Experience bias is possible, as researchers with more experience might have a better understanding of complex tools and methods, leading to more positive evaluations. Less experienced researchers might find the tool more challenging. However, our results showed high usability and preference scores with less experienced users, indicating that experience bias was likely minimal. We would expect that this limitation would result in the less experienced participants in Study 2 to require more time to understand and use DriveGroups, potentially skewing performance evaluations. We mitigated this by providing standardized training sessions for all participants to ensure they had a basic understanding of DriveGroups before evaluation, reducing the learning curve impact.

Finally, this paper proposed a design that helps manage access control by grouping research collaborator roles within the team but does not address issues relating to team structure, data characteristics, and data organization. These are important areas of research but too large to tackle in this single work. Future work may include addressing organizational challenges related to team structure and developing systems to mitigate challenges specific to data characteristics, such as size. This includes further development of usable large file-sharing systems and supporting the dissemination of scientific metadata necessary for proper interpretation of results.

7 Conclusion

Generating and sharing large quantities of data within collaborations are a fundamental part of research. Here we present the results of research revealing that principal investigators struggle to

manage internal and external access to data within projects that span multiple types of collaborators. We further present the design and implementation of a system called DriveGroups, which aims to simplify the data sharing process. Results from our laboratory study indicate that DriveGroups made great strides in terms of usability when compared to unmodified Google Drive, and moderate success in terms of access control, and transparency. By improving the data sharing process, our findings will aid scientists in advancing high-impact research by assisting with collaboration.

8 Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under CRII Award 2153500. The authors would like to express their gratitude for the support, which has been instrumental in the successful completion of this research. The views and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

This research was conducted entirely by a team of undergraduate researchers under the supervision of the principal investigator, Sarah Morrison-Smith. The authors would like to give special thanks to Julia Chang, Catherine O'Brien, Morgan Zee, Hariti Patel, Yiyun Wang, Violet Shi, Emily Ringel, Perrin Anto, and Nicholas Adair for their valuable contributions to the project. While not included as co-authors, their efforts were instrumental in the success of this work.

A Study I

A.1 Participant Demographics

Table 2. Participant backgrounds from the preliminary study.

ID	Research Area	Title	Location
P029	Epidemiology	Faculty	Canada
P033	Animal Sciences	Faculty	United States
P045	Human Computer Interaction and Mechanical Engineering	Faculty	United States
P105	Epidemiology	Post-Doc	United States
P108	Microbiology	Faculty	Australia
P139	Psychology	Faculty	United States
P142	Plant biology	Faculty	United States
P170	Epidemiology	Faculty	United States
P174	Computer Science and Bioinformatics	Research Scientist	United States
P180	Biology	Faculty	United States
P191	Animal Sciences	PhD Student	United States
P192	Human Computer Interaction and Health Informatics	PhD Student	United States
P193	Immunology, Microbiology, and Bioinformatics	Post-Doc	United States
P202	Industrial Hygiene	Post-Doc	United States
P207	Computer Science	Faculty	United States
P223	Veterinarian	Faculty	United States
P254	Biology	Faculty	United States
P269	Chemistry	PhD Student	United States
P309	Epidemiology	Faculty	United States
P396	Proteomics and Metabolomics	Lab Director	United States
P447	Human Computer Interaction	Faculty	United States
P447	Evolutionary Genomics	Faculty	United States
P556	Plant biology	Faculty	United States
P560	Psychology	Faculty	United States
P561	Plant biology	Post-doc	China

Continued on next page

Table 2 – continued from previous page

ID	Research Area	Title	Location
P657	Animal Sciences	Faculty	United States
P678	Psychology	Faculty	United States
P686	Animal Sciences	Faculty	United States
P703	Biology	Faculty	United States
P733	Neurology	Post-Doc	United States
P766	Biology	Faculty	United States
P678	Biology	Faculty	United States
P810	Computer Science	Faculty	United States
P860	Biology	Lab Technician	United States
P897	Biology	Post-Doc	United States
P927	Paleontology	Faculty	United States
P974	Genomic Medicine	Faculty	United States
P982	Computer Science	Faculty	United States

A.2 Semi-Structured Interview Questions for Preliminary Study

A.2.1 Initial Exploration.

- Can you tell us about what you do and what you are currently working on?
- What are the goals and expected outputs for your project?
- What is the potential impact of your work?
- Do you currently share data on your project and how do you do it?
- Which software systems or tools do you use to share data?
- What is easy/difficult about this process?
- What kind of data do you share and with who? (any collaborators - organizations, institutions, etc.)
- How do you control who has access to which data?

A.2.2 Observation.

- Can you walk me through the software you currently use to share data with others?
- If you're comfortable, could you share your screen and talk through your thought process in each decision you make when using the software.
- What works well? Any challenges?
- Can you tell us about the collaboration tools you currently use?
- Walk me through the process you have for sharing/analyzing data in a collaborative team. What part is individual and what requires teamwork?
- What works well? Any challenges? Any collaborators you are uncomfortable sharing data with? Who are those collaborators?
- How do you currently organize your data? What tools do you use?
- Who manages the data? Who has access?

- What works well? Can you tell us about any challenges you face when managing data?
- Is there any data you are uncomfortable sharing with certain collaborators?
- For each tool (you mentioned xyz software tools you use to share/collaborate/organize data)
- Why do you use that tool? / What are the benefits to using this tool over alternatives?
- (Depending on the task) Do you use this tool with all of the collaborators that you do this task with?
- What access control concerns do you currently have and how would you address them?
- What access control concerns do you face with X tool? Can you tell me more about those issues with X tool? Why do you have those issues? ... Y tool? Z tool?
- What happened last time you shared data with others?
- Can you tell me about a time where you were sharing data and ran into a challenge?

A.2.3 Reflection.

- What would your ideal software system do to ensure secure data sharing?
- What would be your ideal way to collectively manage data?
- What functionalities would you prefer to see embedded for more effective collaboration?
- How would you like to give your collaborators access permissions to data? How would you like to receive access permissions?
- If you could create a hypothetical future technology to make collaborating easier, what would it be?

B Study II

B.1 Participant Demographics for Study II

B.2 Usability and Access Control Survey Questions

B.2.1 System Usability Scale (SUS). Scored from Strongly Agree to Strongly Disagree as outlined in Brooke's paper [9]:

- I think that I would like to use this system frequently.
- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I think that I would need the support of a technical person to be able to use this system.
- I found the various functions in this system were well integrated.
- I thought there was too much inconsistency in this system.
- I would imagine that most people would learn to use this system very quickly.
- I found the system very cumbersome to use.
- I felt very confident using the system.
- I needed to learn a lot of things before I could get going with this system.

B.2.2 Other Survey Questions. Scored from 0 to 10 using a visual analog scale, where 0 is Strongly Disagree and 10 is Strongly Agree:

- I think the system allows me to restrict the access to some of my information to some people.
- I think I have control over what information is shared by the system with other people.
- It is clear whether my information is shared with other individuals or groups.
- I believe that the system will prevent unauthorized people from accessing my information.
- I believe my information is accessible only to those authorized to have access.
- I think that there would be a high potential for privacy loss associated with giving my information to the system.
- Considering the information I provide to the system, and the people who might see it, I think it would be risky to give my information to the system.

Table 3. Participant backgrounds from the usability study.

PID	Major(s)	Years Using Google Drive
P055	Environmental Studies & Psychology	8
P072	Data Science	5
P113	Biology & Environmental Studies	1
P119	Chemistry	4
P250	Physics & Mathematics	5
P302	Neuroscience, Mathematics	7
P320	Sociology	8
P352	Computer Science & Mathematics	4
P385	Computer Science & Mathematics	6
P594	Economics & Psychology	2
P603	Hispanic Studies	4
P614	Psychology & Philosophy	2
P652	Computer Science & Mathematics	5
P678	Biochemistry	7
P689	Chemistry	9
P764	Environmental Studies	10
P746	Biology	6
P836	Geosciences	7

- Considering the information I provide to the system, and the people who might see it, I think there is a high potential for information to be shared with the wrong individuals.
- I can understand whether people who I may know (friends, family, classmates, colleagues, acquaintances, etc.) have access to my information on the system.
- It is clear who is the audience of my shared information on the system.
- I am confident I can restrict un-intended people from viewing my information on the system.
- I am confident I can manage who can view my information on the system.
- I think the system is transparent with who my data is shared with.
- I am aware of the system's access control capabilities.
- I am aware of what data is shared with which individuals.

B.3 Semi-Structured Interview Questions for Study II

- Can you describe your initial impressions of DriveGroups/Google Drive?
- Did you encounter any challenges or difficulties while using DriveGroups? What about while using Google Drive?
- Can you describe any features in DriveGroups that you found particularly useful? What about in Google Drive?
- Were there any features that you found unnecessary or confusing?

- Have you encountered any data sharing concerns or issues while using DriveGroups/Google Drive?
- Can you describe any data sharing and access features in DriveGroups/Google Drive that you found particularly reassuring?
- Are there any data control measures you wish DriveGroups/Google Drive would have?
- How transparent is the data access and management process in DriveGroups/Google Drive?
- How likely are you to recommend DriveGroups to a colleague or a friend? Why?
- Are there any features or capabilities you wish DriveGroups would have?
- Is there anything else you would like to share about your experience with DriveGroups
- Are there any features or capabilities you wish DriveGroups would have?
- Is there anything else you would like to share about your experience with DriveGroups?

References

[1] Daniel Atkins. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* Technical Report. Natural Science Foundation.

[2] Carl Auerbach and Louise B Silverstein. 2003. *Qualitative Data: An Introduction to Coding and Analysis.* Vol. 21. NYU press, New York, NY, USA.

[3] Hugh Beyer and Karen Holtzblatt. 1998. *Contextual Design: Defining Customer-centered Systems.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[4] Donna Bonde. 2013. *Qualitative Interviews: When Enough Is Enough.* Technical Report. Research by Design.

[5] Christine L Borgman. 2010. *Research Data: Who Will Share What, with Whom, When, and Why?* Technical Report. RatSWD Working Paper.

[6] Christine L Borgman. 2012. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology* 63, 6 (2012), 1059–1078.

[7] Christine L Borgman, Milena S Golshan, Ashley E Sands, Jillian C Wallis, Rebekah L Cummings, Peter T Darch, and Bernadette M Randles. 2016. Data Management in the Long Tail: Science, Software, and Service. *International Journal of Digital Curation* 11, 1 (2016), 128–149.

[8] Christine L Borgman, Jillian C Wallis, and Matthew S Mayernik. 2012. Who's Got the Data? Interdependencies in Science and Technology Collaborations. *Computer Supported Cooperative Work (CSCW)* 21, 6 (2012), 485–523.

[9] John Brooke. 1995. SUS: A Quick and Dirty Usability Scale. *Usability Eval. Ind.* 189 (Nov. 1995), 4–7.

[10] David L Buckeridge, Robin Mason, Ann Robertson, John Frank, Richard Glazier, Lorraine Purdon, Carl G Amrhein, Nita Chaudhuri, Esme Fuller-Thomson, Peter Gozdyra, David Hulchanski, Byron Moldofsky, Maureen Thompson, and Robert Wright. 2002. Making health data maps: a case study of a community/university research collaboration. *Social Science & Medicine* 55, 7 (2002), 1189–1206. https://doi.org/10.1016/S0277-9536(01)00246-5

[11] Asma Cherif. 2012. *Access Control Models for Collaborative Applications.* Ph. D. Dissertation. Université de Lorraine.

[12] Jean-Paul Chretien, Caitlin M Rivers, and Michael A Johansson. 2016. Make Data Sharing Routine to Prepare for Public Health Emergencies. *PLoS medicine* 13, 8 (2016), e1002109.

[13] Mick P Couper, Roger Tourangeau, Frederick G Conrad, and Eleanor Singer. 2006. Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment. *Social Science Computer Review* 24, 2 (2006), 227–245.

[14] Melissa H Cragin, Carole L Palmer, Jacob R Carlson, and Michael Witt. 2010. Data Sharing, Small Science and Institutional Repositories. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 368, 1926 (2010), 4023–4038.

[15] Robert Darby, Simon Lambert, Brian Matthews, Michael Wilson, Kathrin Gitmans, Suenje Dallmeier-Tiessen, Salvatore Mele, and Jari Suhonen. 2012. Enabling Scientific Data Sharing and Re-Use. In *2012 IEEE 8th International Conference on E-Science.* IEEE, Washington, DC, USA, 1–8. https://doi.org/10.1109/eScience.2012.6404476

[16] Devan Ray Donaldson and Joshua Wolfgang Koepke. 2022. A Focus Groups Study on Data Sharing and Research Data Management. *Scientific Data* 9, 1 (2022), 345.

[17] Dropbox Inc. 2021. Dropbox.

[18] Clifford S Duke and John H Porter. 2013. The Ethics of Data Sharing and Reuse in Biology. *BioScience* 63, 6 (2013), 483–489.

[19] Paul Edwards, Matthew Mayernik, Archer Batcheller, Geoffrey Bowker, and Christine Borgman. 2011. Science Friction: Data, Metadata, and Collaboration. *Social studies of science* 41 (10 2011), 667–90. https://doi.org/10.2307/41301955

[20] Ixchel M Faniel and Trond E Jacobsen. 2010. Reusing Scientific Data: A Research Framework. *Ann Arbor* 1001 (2010), 48109–1107.

[21] Benedikt Fecher, Sascha Friesike, and Marcel Hebing. 2015. What Drives Academic Data Sharing? *PloS one* 10, 2 (2015), e0118053.

[22] Ana Sofia Figueiredo. 2017. Data Sharing: Convert Challenges into Opportunities. *Frontiers in public health* 5 (2017), 327.

[23] Bobby Lee Houtkoop, Chris Chambers, Malcolm Macleod, Dorothy VM Bishop, Thomas E Nichols, and Eric-Jan Wagenmakers. 2018. Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in methods and practices in psychological science* 1, 1 (2018), 70–85.

[24] Andreas Hueni, Jens Nieke, Juerg Schopfer, Mathias Kneubühler, and Klaus I Itten. 2009. The Spectral Database SPECCHIO for Improved Long-Term Usability and Data Sharing. *Computers & Geosciences* 35, 3 (2009), 557–565. https://doi.org/10.1016/j.cageo.2008.03.015

[25] Claudia-Lavinia Ignat. 2021. *Large-Scale Trustworthy Distributed Collaboration.* Authorization \ ‘a to Direct Research. University \ ’e of Lorraine.

[26] Susan Ivey, Sophia Lafferty-Hess, Peace Ossom-Williamson, and Katie Barrick. 2023. Managing, Sharing, and Publishing Data. In *Scholarly Communication Librarianship and Open Knowledge.* Association of College and Research Libraries, Chicago, Illinois.

[27] Jing Jin and Gail-Joon Ahn. 2006. Role-Based Access Management for Ad-Hoc Collaborative Sharing. In *Proceedings of the Eleventh ACM Symposium on Access Control Models and Technologies (SACMAT ’06).* Association for Computing Machinery, New York, NY, USA, 200–209. https://doi.org/10.1145/1133058.1133086

[28] Maritza Johnson, Steven Bellovin, Robert Reeder, and Stuart Schechter. 2009. Laissez-Faire File Sharing Access Control Designed for Individuals at the Endpoints. In *Proceedings New Security Paradigms Workshop.* ACM, Oxford, United Kingdom, 1–10. https://doi.org/10.1145/1719030.1719032

[29] Stefanie K Johnson, Kenneth Bettenhausen, and Ellie Gibbons. 2009. Realities of Working in Virtual Teams: Affective and Attitudinal Outcomes of Using Computer-Mediated Communication. *Small Group Research* 40, 6 (2009), 623–649.

[30] Eunice Jun, Blue A. Jo, Nigini Oliveira, and Katharina Reinecke. 2018. Digestif: Promoting Science Communication in Online Experiments. In *Proc. of the ACM on Human-Computer Interaction*, Vol. 2 (CSCW). ACM, New York, NY, United States, 1–26.

[31] Youngseek Kim, Benjamin K Addom, and Jeffrey M Stanton. 2011. Education for eScience Professionals: Integrating Data Curation and Cyberinfrastructure. *International journal of digital curation* 6, 1 (2011), 125–138.

[32] Laura Koesten, Emilia Kacprzak, Jeni Tennison, and Elena Simperl. 2019. Collaborative Practices with Structured Data: Do Tools Support What Users Need?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’19).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300330

[33] Ben Langmead and Abhinav Nellore. 2018. Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics* 19 (01 2018). https://doi.org/10.1038/nrg.2017.113

[34] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2010. *Research Methods in Human-Computer Interaction.* Wiley Publishing, Hoboken, NJ, United States.

[35] Rahuman S Malik-Sheriff, Mihai Glont, Tung V N Nguyen, Krishna Tiwari, Matthew G Roberts, Ashley Xavier, Manh T Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, Emma L Fairbanks, Johannes P Meyer, Chinmay Arankalle, Thawfeek M Varusai, Vincent Knight-Schrijver, Lu Li, Corina Dueñas-Roca, Gaurhari Dass, Sarah M Keating, Young M Park, Nicola Buso, Nicolas Rodriguez, Michael Hucka, and Henning Hermjakob. 2020. BioModels—15 Years of Sharing Computational Models in Life Science. *Nucleic Acids Research* 48, D1 (2020), D407–D415.

[36] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R. Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? *Proc. ACM Hum.-Comput. Interact.* 3, GROUP (Dec. 2019), 237. https://doi.org/10.1145/3361118

[37] Roy A. Maxion and Robert W. Reeder. 2005. Improving User-Interface Dependability through Mitigation of Human Error. *International Journal of human-computer studies* 63, 1-2 (2005), 25–50.

[38] William K. Michener. 2015. Ecological Data Sharing. *Ecological Informatics* 29 (2015), 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010

[39] Sarah Morrison-Smith, Christina Boucher, Andrea Bunt, and Jaime Ruiz. 2015. Elucidating the Role and Use of Bioinformatics Software in Life Science Research. In *Proceedings of the 2015 British HCI Conference.* ACM, Lincoln, Lincolnshire, United Kingdom, 230–238.

[40] Gary M. Olson and Judith S. Olson. 2000. Distance Matters. *Human-Computer Interaction* 15, 2 (Sept. 2000), 139–178. https://doi.org/10.1207/S15327051HCI1523_4

[41] Judith S Olson and Gary M Olson. 2006. Bridging Distance: Empirical Studies of Distributed Teams. *Human-Computer Interaction in Management Information Systems* 2 (2006), 27–30.

[42] Reddit.com. 2023. AskScience: Got Questions? Get Answers.

[43] Diana K Smetters and Nathan Good. 2009. How Users Use Access Control. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, Mountain View, California, USA, 15.

[44] H. Subramonyam, S.M. Drucker, and E. Adar. 2019. Affinity Lens: Data-Assisted Affinity Diagramming with Augmented Reality. In *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, United States, 1–13.

[45] Subrahmaniam Tangirala and Bradley J Alge. 2006. Reactions to Unfair Events in Computer-Mediated Groups: A Test of Uncertainty Management Theory. *Organizational behavior and human decision processes* 100, 1 (2006), 1–20.

[46] Leho Tedersoo, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, Marju Raju, Anastasiya Astapova, Heli Lukner, Karin Kogermann, and Tuul Sepp. 2021. Data Sharing Practices and Data Availability upon Request Differ across Scientific Disciplines. *Scientific Data* 8, 1 (July 2021), 192. https://doi.org/10.1038/s41597-021-00981-0

[47] Anand Tripathi, Tanvir Ahmed, Richa Kumar, and Shremattie Jaman. 2020. A Coordination Model for Secure Collaboration. In *Process Coordination and Ubiquitous Computing*. CRC Press, Boca Raton, FL, United States, 77–95.

[48] Anand R Tripathi, Tanvir Ahmed, and Richa Kumar. 2003. Specification of Secure Distributed Collaboration Systems. In *The Sixth International Symposium on Autonomous Decentralized Systems, 2003. ISADS 2003.* IEEE, New York, NY, United States, 149–156.

[49] Nancy A Van House, Mark H Butler, and Lisa R Schiff. 1998. Cooperative Knowledge Work and Practices of Trust: Sharing Environmental Planning Data Sets. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*. ACM, Seattle, Washington, USA, 335–343.

[50] Theresa Velden. 2013. Explaining Field Differences in Openness and Sharing in Scientific Communities. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, San Antonio, Texas, USA, 445–458.

[51] Janet Vertesi and Paul Dourish. 2011. The Value of Data: Considering the Context of Production in Data Economies. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 533–542. https://doi.org/10.1145/1958824.1958906

[52] Jillian C Wallis, Elizabeth Rolando, and Christine L Borgman. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PloS one* 8, 7 (2013), e67332.

[53] Tara Whalen, Diana Smetters, and Elizabeth F. Churchill. 2006. User Experiences with Sharing and Access Control. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*. ACM, Montréal, Canada, 1517–1522.

[54] Daniel Woodraska, Michael Sanford, and Dianxiang Xu. 2011. Security Mutation Testing of the FileZilla FTP Server. In *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC '11)*. Association for Computing Machinery, New York, NY, USA, 1425–1430. https://doi.org/10.1145/1982185.1982493

[55] Liu Xia, Feng Chao-sheng, Yuan Ding, and Wang Can. 2010. Design of Secure FTP System. In *2010 International Conference on Communications, Circuits and Systems (ICCCAS)*. IEEE, New York, NY, United States, 270–273. https://doi.org/10.1109/ICCCAS.2010.5582002

[56] Tianyin Xu, Han Min Naing, Le Lu, and Yuanyuan Zhou. 2017. How Do System Administrators Resolve Access-Denied Issues in the Real World?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver, Colorado, USA, 348–361.

Received May 2024; revised August 2024; accepted October 2024