Personal Data Curation in the Cloud Age

Individual Differences and Design Opportunities

by

Francesco Vitale

B.Sc., University of Milano-Bicocca, 2013

M.Sc., KTH Royal Institute of Technology, 2016

M.Sc. Mention très bien, Université Paris-Saclay, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL
STUDIES

(Computer Science)

The University of British Columbia
(Vancouver)

July 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Personal Data Curation in the Cloud Age: Individual Differences and Design Opportunities

submitted by Francesco Vitale in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Computer Science

Examining Committee:

Joanna McGrenere, Professor, Department of Computer Science, UBC
Supervisor

William Odom, Assistant Professor, School of Interactive Arts & Technology, Simon Fraser University
Supervisory Committee Member

Ivan Beschastnikh, Associate Professor, Department of Computer Science, UBC
Supervisory Committee Member

Margo Seltzer, Professor, Department of Computer Science, UBC
University Examiner

Heather O'Brien, Associate Professor, iSchool, UBC
University Examiner

Additional Supervisory Committee Members:

Karon MacLean, Professor, Department of Computer Science, UBC
Supervisory Committee Member

Abstract

People are creating and storing a growing amount of personal data, from photos and documents, to messages and applications, on a growing number of devices. Storage space, often in the cloud, is cheap or free. But previous research shows that a degree of selectivity and curation is necessary to build personal archives that have value over time.

In this dissertation, we ask: How do different people decide what personal data to keep or discard? What drives their decisions? And how can data management tools better support individual preferences?

We used a qualitative and design-based approach to conduct four studies consisting of 64 interviews in total and a survey (n=349).

First, we identified a spectrum of tendencies that informed how participants (n=23) decided what to keep or discard, with two extremes: "hoarding" (keeping most of data), and "minimalism" (keeping as little as possible). We extended this spectrum with a set of five behavioral styles that capture contextual curation patterns: taking a casual approach to data, feeling overwhelmed, collecting data, purging data, and trying to be frugal. This model of behaviors (based on the 64 interviews) highlights a key role for data curation: what people keep or discard informs how they think about their own identity.

We used these insights to map a design space for data curation and create five design concepts for different user needs, exploring automation and other key design dimensions. Participants' reactions (n=16) varied: some welcomed technology and automation, others opposed it, with context informing their reactions.

Inspired by these results and using a taxonomy of data types and decluttering criteria based on the survey (n=349), we designed Data Dashboard, a tool that

aggregates data from a user's multitude of devices and cloud platforms, providing customizable functions for different goals. We evaluated a prototype of the system with 18 participants and found that a personalized approach to data curation is promising, so long as it respects users' boundaries.

Our work outlines key design directions and opportunities that can help envision new tools, prioritize user needs, and redefine our relationship with personal data in a world full of it.

Lay Summary

As people create, save, and share a growing amount of personal digital data (for example, photos, documents, and mobile apps) on their devices and online platforms, how do they decide what to keep or discard? To find out, we ran 64 interviews and an online survey (349 participants). We identified how some participants tended to keep almost everything, while others tried to discard as much as possible. Participants often took a varied approach based on different categories of items (for example, keeping all photos but discarding most apps). Using these results, we created five design concepts and a prototype data management tool that could help users decide what to keep or discard. We evaluated the concepts and the tool with participants and found that because people's preferences differ, a good approach is to create tools that can be personalized to individual desires. Our work provides directions to build such tools.

Preface

Research is a collaborative effort. All the studies I present in this dissertation involved a set of collaborators in different capacities.

My supervisor Dr. Joanna McGrenere followed my PhD work from start to finish. After Study 1 (Chapter 4), Dr. William Odom joined as a member of my PhD committee and an unofficial co-supervisor. Dr. McGrenere and Dr. Odom jointly supervised and helped with all the additional studies that make up my PhD (Chapter 5, Chapter 6, Chapter 7, Chapter 8). Here, I list the collaborators involved in each study and detail my role and contributions.

In **Chapter 4**, I report Study 1, that I conducted as part of my RPE (Research Proficiency Evaluation) [1] from May to August 2017. I designed the study, choosing data collection and analysis methods, guided by my supervisor Dr. Joanna McGrenere. I conducted and transcribed all 23 interviews. I analyzed the data helped by PhD student Izabelle Janzen and my supervisor Dr. Joanna McGrenere, as described in the chapter. I was the lead author of the final report and paper based on it, later published at the ACM CHI 2018 conference [290], where it won a Best Paper Award (top 1% of all paper submissions). Because the paper has been peer-reviewed, in Chapter 4, I reproduce its contents with a few adjustments:

> **Francesco Vitale**, Izabelle Janzen, and Joanna McGrenere. (2018) *Hoarding and Minimalism: Tendencies in Digital Data Preservation*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18).

[1] The RPE process consists of a four-month research project, followed by an examination from an RPE supervisory committee. More details about the process are available online: https://www.cs.ubc.ca/students/grad/graduate-programs/research-proficiency-evaluation-rpe.

In **Chapter 5**, I report Study 2, consisting of 7 interviews and an online survey. This study took place in the first half of 2018. I designed the interview study and the survey with the guidance of my co-supervisors Dr. Joanna McGrenere and Dr. William Odom. I recruited participants and then collected and analysed all data, discussing it with my co-supervisors.

In **Chapter 6**, I report an overarching analysis of the four main studies in my PhD, introducing a set of five behavioral styles or archetypes. I originally had the idea of developing a set of "archetypes" in the summer of 2018, after conducting the first two interview studies and the survey. As the chapter explains, I iterated on the analysis throughout my PhD helped by my co-supervisors Dr. Joanna McGrenere and Dr. William Odom, but I am largely responsible for the analysis process. My co-supervisors helped in structuring the chapter and its contributions. My supervisory committee also helped in framing the structure of this chapter.

In **Chapter 7**, I report Study 3, a design study that took place in the second half of 2018. I developed the design concepts and defined the study protocol with the guidance of my co-supervisors Dr. William Odom and Dr. Joanna McGrenere. In particular, Dr. Odom suggested the specific methods we used, centred around creating different design concepts and video prototypes. I recruited, conducted, and analysed all of the interviews in the elicitation study. I wrote a full paper based on the study, with input from my co-supervisors Dr. Odom and Dr. McGrenere. The paper was published at the ACM DIS 2019 conference [291], where it received a Best Paper Honorable Mention (top 2% of all paper submissions). Because the paper has been peer-reviewed, in Chapter 7, I reproduce its contents with some adjustments and additions, based on comments from my committee:

> **Francesco Vitale**, William Odom, and Joanna McGrenere. (2019) *Keeping and Discarding Personal Data: Exploring a Design Space*. Proceedings of the 2019 Conference on Designing Interactive Systems (DIS '19).

In **Chapter 8**, I report Study 4, a design study that took place at the end of 2019 (after I spent the summer and early fall of 2019 at Google as a User Experience Research Intern). I designed and developed the bulk of the Data Dashboard prototype in the first half of 2019, before my internship. From September 2019, undergraduate student Janet Chen joined the project to help finish implementing some functions of the prototype. Janet Chen also assisted in running the evaluation study. With input from my co-supervisors Dr. Joanna McGrenere and Dr. William Odom, I decided the methods to use and designed the study protocol. I recruited all participants and moderated all sessions and pilots, except for the last two interview sessions, that were moderated by Janet Chen. Janet Chen also acted as a note-taker in several interview sessions, helped with data analysis, conducted a literature review of past design work in Personal Information Management (PIM) under my guidance, and contributed to writing the Data Dashboard usability report (Appendix A). I wrote a full paper based on the study with input from Janet Chen, and my co-supervisors Dr. William Odom and Dr. Joanna McGrenere. The paper was published at the ACM DIS 2020 conference [292]. Because the paper has been peer-reviewed, in Chapter 8, I reproduce its contents with some adjustements and additions, based on comments from my committee:

> **Francesco Vitale**, Janet Chen, William Odom, and Joanna McGrenere. (2020) *Data Dashboard: Exploring Centralization and Customization in Personal Data Curation*. Proceedings of the 2020 Conference on Designing Interactive Systems (DIS '20).

Content from the three papers published at CHI and DIS also appears in Chapter 1 and Chapter 2. These chapters also include revised content from an extended abstract accepted to the Doctoral Consortium (a curated track) of the ACM CHI 2019 conference [289]:

> **Francesco Vitale**. (2019) *Designing for Long-term Digital Data Management*. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts).

In the bulk of the dissertation, I write about my work using the plural first-person to acknowledge the collaborative nature of the research studies that I present. I use the first person in Chapter 3 to reflect on my research positionality and take ownership of the research process as a whole.

The UBC Behavioural Research Ethics Board (BREB) approved all studies reported in this dissertation: certificate number H17-00734.

A note on terminology

Throughout the dissertation, we use different expressions to indicate specific user actions that we are interested in studying: *data preservation* (Chapter 4), *decluttering* (Chapter 5), *data curation* (Chapter 6 and Chapter 8), *data selection* (Chapter 7), *keeping and discarding decisions* (throughout the whole dissertation). These expressions refer to very similar notions and reflect the evolution of our investigation. We use them on a chapter basis because they reflect our thinking in each specific phase of the research. In Chapter 2, we better explain differences, overlaps, and nuances in meaning. We also provide a definition of each term on a chapter basis.

Table of Contents

List of Tables

List of Figures

List of Supplementary Material

1. Video prototype for the Patina concept from Study 3 (`study3-patina.mp4`)
2. Video prototype for the Data Recommender concept from Study 3 (`study3-recommender.mp4`)
3. Video prototype for the Temporary Folder concept from Study 3 (`study3-temp-folder.mp4`)
4. Video prototype for the Temporary App concept from Study 3 (`study3-temp-app.mp4`)
5. Video prototype for the Future Filters concept from Study 3 (`study3-filters.mp4`)
6. Video walkthrough of the Data Dashboard prototype from Study 4 (`study4-dashboard.mp4`)

Glossary

FM File Management

HCI Human-Computer Interaction

PIM Personal Information Management

UI User Interface

Acknowledgments

I want to start by thanking my main co-supervisor Dr. Joanna McGrenere, for her constant guidance and support over the years. Patient, encouraging, enthusiastic, and always open to new ideas: I could have not asked for a better supervisor.

Next, I want to thank my co-supervisor Dr. William Odom, from whom I have learned a lot about using design methods to drive a research inquiry. His guidance was essential for shaping much of my work.

I am grateful to my supervisory committee members, Dr. Karon MacLean and Dr. Ivan Beschastnikh, for their comments over the years. In particular, I thank Dr. MacLean for challenging key aspects of my work and pushing for better clarity in describing my ideas. I thank Dr. Beschastnikh for providing a much needed technical perspective to my work, with engaging discussions around user behaviors and file systems.

I am also thankful to Dr. Anna Cox, external examiner, and Dr. Margo Seltzer and Dr. Heather O'Brien, university examiners, for their thoughtful comments, questions, and input on my work. Their on-point suggestions helped me improve key parts of the dissertation.

I also want to thank Dr. Sandra Mathison and Dr. Marla Buchanan, instructors of UBC's 2017 winter edition of EPSE 595, for introducing me to key works on qualitative research that I have used in my studies.

A mention goes to all UBC's MUX lab members (students, post-docs, faculty) between 2016 and 2020 who have provided feedback on my work or who have been around to chat during these years. Izabelle Janzen, Oliver Schneider, Soheil Kianzad, Kamyar Ardekani, Laura Cang, Paul Bucci, Yelim Kim, Taslim Arefin Khan, Shareen Mahmud, Sabrina Hauser, Michael Opperman, Jessalyn Alvina, Emmanouil Giannisakis, Hanieh Shakeri, Ashish Chopra, Kevin Chow, Sang-Wha Sien, Frances Sin, Hayley Guillou, Mohi Reza, Hasti Seifi, Matthew Chun, Dilan Ustek, Mona Haraty, Helen Yin, Zack Wilson, Anna Maria Offenwanger, Dongwook Yoon, Tamara Munzner, Leila Aflatoony. A special thanks to Janet Chen who helped me during the Data Dashboard study and who has been a great student to supervise.

I also want to thank the team members of my group project in UBC's 2017 edition of CPSC 554m: Matthew Chun, Yingying Wang, Haotian Zhang. Our work on collaborative file management inspired a small feature of the Data Dashboard prototype in Chapter 8.

Several people at UBC have also helped me in different roles over the years. I thank Dr. Alla Sheffer for chairing my research proficiency evaluation presentation, Dr. Ronald Garcia for chairing my thesis proposal defense, Dr. Eric Meyers for chairing my PhD defense, Laura Selander and Joyce Poon for their administrative support.

I thank the organizers, mentors, and fellow participants at the CHI 2019 Doctoral Consortium for their feedback and suggestions on my thesis: Laura Dabbish, Koji Yatani, Duncan Brumby, Jennifer Thom, Walter Lasecki, Florian 'Floyd' Mueller, Elena Agapie, Alex Ahmed, Yulia Brazauskayte, Simran Chopra, Karen Cochrane, Michael Ann DeVito, Jessica Feuston, Jane Hoffswell, Suvi Holm, Matthew Hong, Tom Horak, Fannie Liu, Tricia Ngoon, Makuochi Nkwo, Teddy Seyed, Rachel Simons, David Verweij, Jillian Warren, Cara Wilson.

I also want to thank the several anonymous reviewers at CHI, DIS, and CSCW who have taken the time to provide feedback on my submissions over the years. Their comments have contributed in improving my work.

Many people that I have chatted with at CHI, DIS, or in other occasions have also helped my work indirectly. Among others: Jofish Kaye, Steve Whittaker, Anne Spaa, Jesse David Dinneen, Ofer Bergman.

A thanks also goes to my team and collaborators at Google, from whom I have learned several research tips and techniques that I have used in my last study and beyond: Melinda Klayman, Kerstin Kuehne, Caroline Nivetha.

In Study 3 (Chapter 7) we created the concepts using Sketch and InVision with icons from the Streamline library and The Noun Project (by Abdo, AR Ehsan, Dino, Hare Krishna, and Landan Lloyd). The prototype built in Study 4 (Chapter 8) includes icons from Flat Icon.

All my research, of course, would not have been possible without the many participants in my studies: I thank them for taking them time to help me understand how they curate their data and I hope that my work might help solve some of their problems in the future.

Finally, I am thankful for the support of my family and friends, on one or the other side of the world.

My PhD was supported by the grant NSERC RGPIN-2017-04549 "Highly personalized user interfaces."

Chapter 1

Introduction

In this chapter, we provide an overview of the dissertation, starting from context and motivation[1]. Then, we outline our research questions, goals, and approach. We introduce our contribution and then conclude with an outline of the dissertation.

1.1 Context and motivation

Why study personal data curation? We argue that the current technological landscape calls for a renewed interest and attention to how people manage their personal data. Here, we explain why.

1.1.1 A seductive digital landscape

The past decade has changed the way people think about personal data. Ten years ago cloud storage platforms were in their infancy. Dropbox launched in 2008, iCloud in 2011, Google Drive in 2012. Before the advent of these platforms and before the popularity of mobile devices, people largely thought of their data as limited to files and folders stored on local computers. Now, personal data is a buzzword at the center of political debates and regulatory efforts [1, 265]. Data dominates all aspects of life, from work life to domestic life, so much so that an entire industry relies on the storing, accumulation, and exchange of personal data

[1] Part of the content in this chapter also appears in the publications mentioned in the Preface to the dissertation.

for profit [50, 264, 319]. This is a "seductive digital landscape" [183] where trends such as lifelogging technologies [85] and cheap storage space [155] allow people to collect and keep a virtually infinite amount of data. It is not surprising that collections of personal digital items are growing [73, 76, 213, 285].

But these technological trends can lead to questions about the role of data in daily life. How can people extract value and meaning from large quantities of data over time? [133, 157, 229] How can permanent records of personal data allow for forgetting? [11, 254] And how can individual behaviors scale to a growing population? [84, 236, 285]

1.1.2 The need to focus on personal data *curation*

Researchers in Personal Information Management (PIM) have called for an increased focus on studying and designing for long-term data management practices [144]. Key research questions to address include how to enable meaningful data legacy [111], how to approach automation of data management [144], and how to support data *curation* [302]. In this dissertation, we focus on *personal data curation*.

Personal data *curation* is the process of saving and managing data for later use. It consists of three stages: *keeping* (where users create or save data, and then decide what to keep or discard), *managing* (where users organize kept items), and *exploiting* (where users retrieve and use the information they have kept) [302]. Improvements in search functions have made retrieving items easier. Instead, the keeping and managing phases of curation challenge users. In particular, users struggle when deciding what items to keep or discard, often opting to keep as a default action [38].

1.1.3 Why discarding data is hard

There are three main reasons that make discarding data challenging. First, discarding data can lead to a "paradox" [26]: you might want to delete unneeded data to focus your time and attention on what is important. But deciding what to delete will take time and attention. Second, deciding what to keep or not is hard because it often involves predicting the future, something that humans are not always good at or might find stressful. People might decide to not take any intentional decision and keep everything just to avoid missing any useful data in the future [24, 145]. Third, users often perceive the activity of sorting personal data as burdensome because of the poor support that data management tools provide. These challenges often create an unintentional accumulation cycle: the more you keep, the harder it becomes to curate [24], with people becoming increasingly less aware of what they have in the first place [213].

More and more, research studies report how technology users can feel stressed about the accumulation of data [171, 205, 253, 271]. Previous studies show how people might need some degree of selection to make sense of their digital archives, both for data they explicitly create or acquire [166, 182, 183] and also for data created by technology on their behalf [308]. There are situations (e.g., romantic breakups) where people assume specific roles in relation to curating data [247]: deleter, keeper, selective disposer. But outside of these specific situations, a common finding is that people often "loathe" deleting data [205, 308]. These insights into people's keeping and discarding practices, however, are broad and do not consider individual differences.

1.1.4 Why look at individual differences

Most of the literature on PIM focuses on how people *organize* or *retrieve* information. One recurring finding is that people show substantial individual differences in their behavior. For example, some people choose to "neatly" manage files with organized structures; others let them sit within "messy" piles [127]. Some keep emails in their inbox; others clean them regularly [117]. Some use search to retrieve a piece of information; others rely on navigation [25, 28]. We see some features of modern data management tools as likely informed by this line of research (e.g., macOS' automatic "Stacks" for creating piles of desktop files [8]). These features show the importance of looking at individual differences within PIM.

However, there is still a lot to understand about the data curation process as a whole. In particular, the literature lacks a deep understanding of behaviors in the initial stage of curation, where people acquire data and then decide what to keep or discard [302]. Studies in Human-Computer Interaction (HCI) and Information Science report how people's digital archiving strategies are "widely varied" [259].

But despite knowing this, there is no systematic model of individual behaviors, a need that many previous studies call out [23, 118, 150, 309]. This is the space that this thesis explores. In particular, **we look at keeping and discarding decisions in personal data curation**.

1.1.5 Why studying personal data curation matters

The need for better curation tools also comes from popular discourse about technology and people's demand for more control over their data. Privacy is a key concern for people [176]. The advent of the Cloud has imposed a centralized data management model where a few monopolizing corporations (Amazon, Apple, Google, Facebook, Microsoft) aggregate people's data. But in recent years the pitfalls of this model have come to light: from government-sanctioned surveillance programs [277] and illicit use of personal data for political aims [283], to photo leaks [9] and voice assistants secretly sharing private conversations [172]. These episodes have made people more aware of their own "digital footprint" and the consequences of an economy increasingly based on data. To companies, data is nothing more than a commodity, no matter how private or personal. But people *want* more control over their data [176]. As the importance of digital data continues to grow, current models face increased scrutiny. This is also why we see a need for better tools to manage and curate personal data.

1.2 Research goals, research questions, and approach

This dissertation has three main goals:

1. Providing a rich, empirical account of user practices in personal data curation, with a focus on keeping and discarding decisions (an under-explored area of curation). To reach this goal, we ask the following research questions: *How do people decide what personal data to keep or discard? What drives their decisions?* (Chapter 4) *What are the types of data that people consider when curating, and what are the criteria and strategies they use to "declutter"?* (Chapter 5)

2. Extending previous characterizations of individual differences in PIM by focusing on the first stage of curation and taking a design-focused perspective. To reach this goal, we ask the following research question (Chapter 6): *How can we make individual differences in personal data curation actionable for research and design?*

3. Exploring and evaluating a range of design approaches for supporting different user needs. To reach this goal, we ask the following research questions: *How can we design technologies to support people's individual decisions around what data to keep or discard? In particular, what different design approaches might be viable to different users and in different situations?* (Chapter 7) *Can centralization and customization help people decide what personal data to keep or discard?* (Chapter 8)

To answer the research questions above, we conducted four studies (Figure 1.1):

- Study 1 (Chapter 4): an exploratory interview study with 23 participants.
- Study 2 (Chapter 5): a mixed-methods study consisting of 7 contextual interviews and an online survey (349 participants).
- Study 3 (Chapter 7): a design exploration with 16 participants.
- Study 4 (Chapter 8): a system evaluation with 18 participants.

The studies combine different methodological approaches: from a grounded theory approach in Study 1 and Study 2, to a *Research through Design* approach in Study 3 and Study 4. In Chapter 3, we provide more details on the overall methodology across studies.

1.3 Contributions

The primary contributions of the dissertation fall under three areas: theoretical, design and artifacts, empirical [310].

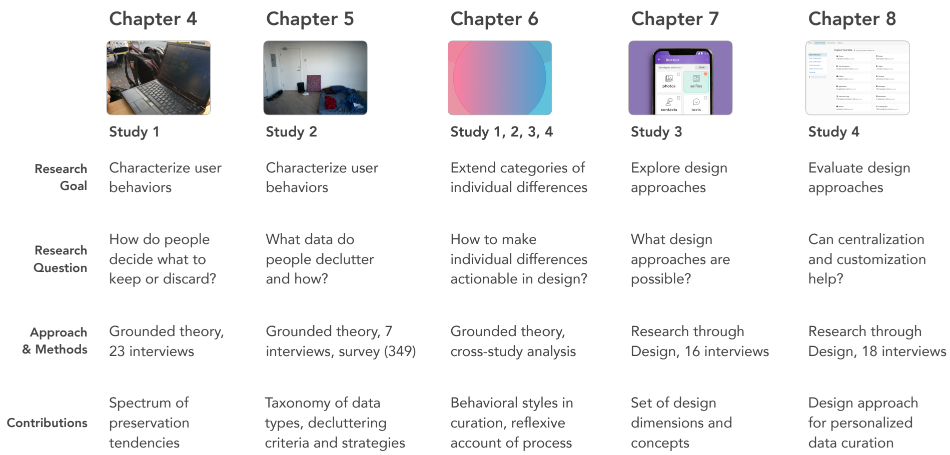


Figure 1.1: An overview of research chapters in the dissertation and the studies reported in each. For each chapter, from top to bottom, we outline: the research goal, the key research question, the research approach and methods, the key contributions.

- *Theoretical contributions*: a rich description of individual differences in personal data curation consisting of a spectrum of general preservation tendencies (Chapter 4) and a set of five behavioral styles (Chapter 6), with an explanation of their role for identity construction over time.

- *Design and artifact contributions*: a set of dimensions and concepts for defining and exploring a design space around personal data curation (Chapter 7), a design approach to address key challenges in data curation by combining a unified tool with personalized functions (Chapter 8), a set of generative design directions and opportunities across all four studies (Chapter 4, Chapter 5, Chapter 6, Chapter 7, Chapter 8).

- *Empirical contributions*: four studies (64 interviews in total and an online survey with 349 participants) that report personal data curation practices with the use of different methods and approaches (Chapter 4, Chapter 5, Chapter 6, Chapter 7, Chapter 8).

In addition, the secondary contributions of the dissertation are: a reflexive account of our research and design process across the four studies (Chapter 6); a taxonomy of personal data types and decluttering criteria (Chapter 5); and a description of temporal decluttering practices (Chapter 5).

1.4 Outline

The rest of the dissertation follows this structure. In Chapter 2, we provide some background and definitions for personal data management and curation, framing our work against two theoretical frameworks that inform our work. We also give an overview of related work on personal data management and outline key insights about people's keeping and discarding practices.

Chapter 3 presents the overall methodological approach of the thesis, with details on methods and criteria for research quality.

In Chapter 4, we present the results from Study 1, where we conducted exploratory interviews with a broad sample of 23 participants. In this study, we introduce the idea of a spectrum of general tendencies for data preservation with two extremes: "hoarding" (keeping most of data) and "minimalism" (trying to keep as little as possible). The spectrum represents a first take on individual differences in personal data curation.

In Chapter 5, we present the results of Study 2, which was a followup study extending the results on "hoarding" and "minimalism" using a mixed-methods approach. First, we conducted contextual interviews with 7 self-identified "minimalists." In parallel, we ran an online survey with 349 participants. We use open-ended survey responses to build a taxonomy of data types with related "decluttering" criteria. We later used these results to inform our design work in the last two studies. Then, based on data from both the survey and the interviews, we describe a set of decluttering practices and selection strategies that we later use to inform both our work on user modelling and our design work.

In Chapter 6, we present one of the key contributions of the dissertation: a set of five behavioral styles in personal data curation: *Casual*, *Overwhelmed*, *Collector*, *Purger*, and *Frugal*. To describe the behavioral styles, we draw on all the four interview studies that make up the dissertation. First, we explain how we identified the behavioral styles after conducting Study 1 and Study 2. Then, we bring forward data from the last two studies (described in more detail in Chapter 7 and Chapter 8) to show how we enriched and validated the behavioral styles.

In Chapter 7, we present Study 3, a design study focused on exploring the design space of personal data curation with five concepts that can help users select what data to keep or discard. These are *Patina*, *Data Recommender*, *Temporary Folder*, *Temporary App*, and *Future Filters*. We elicited reactions to the concepts from a varied sample of 16 participants (recruited using a short description of the behavioral styles presented in Chapter 6).

In Chapter 8, we present Study 4, a design study focused on evaluating *Data Dashboard*, a prototype system for curating personal data. The system embodies insights from all previous studies, with some functions taking inspiration from the design concepts and insights presented in Chapter 7, or modeled after the patterns of behaviors from previous chapters. We evaluated the prototype with a varied sample of 18 participants (recruited using the same short description of the behavioral styles from Chapter 6 that we used in the previous design study).

Finally, in Chapter 9, we summarize our results and discuss the implications of our work.

Chapter 2

Background

In this chapter, we give an overview of past studies and general user practices in personal data curation[1]. The next chapters will include more specific related work tied to aspects of each study. In particular, in Chapter 4, we will discuss related work about digital hoarding. In Chapter 6, we will provide more details about individual differences in PIM. In Chapter 8 we will focus on previous work on personalization and customization. In both Chapter 7 and Chapter 8 we will discuss past design and system work in PIM.

2.1 Defining personal data and personal data curation

To define the scope of our work, we first draw some distinctions between "personal data management," PIM and File Management (FM). Then, we introduce two frameworks that guide our work.

2.1.1 Framing personal data management

PIM is a broad activity that involves creating, acquiring, organising, interacting with, and searching for *personal information* [27, 142]. PIM research within the field of HCI goes back to the advent of personal desktop computers [158, 177].

[1]Part of the content in this chapter also appears in the publications mentioned in the Preface to the dissertation.

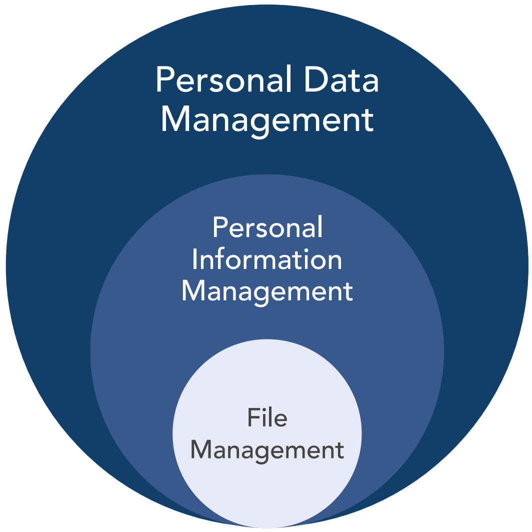


Figure 2.1: We conceptualize "personal data management" as broader than "personal information management," that in turn is broader than "file management." (The size of the circles in this illustration represents how broad the focus of each term is, not their importance.)

Today, the literature and popular discourse largely think of personal information as synonymous with digital information but that was not always the case. Several early PIM studies focus on physical information, such as paper documents, and investigate how to support people's management practices when moving from physical to digital information [158, 177, 303]. As the bulk of people's work and daily life shifted to digital, PIM studies honed in on emerging types of information: computer documents [14, 38], emails [117, 306], contacts [305], bookmarks [273].

Compared to PIM, *File Management* has a narrower focus, limited to files and folders [74]. From a technical standpoint, all personal information resides in files, but files are a debated data structure metaphor [123, 191, 255]. Users might experience visual files and folders as different from information such as contacts or bookmarks, that are more integrated into specific applications.

Recently, HCI studies have started using terms such as "personal data," "digital data," and "digital or virtual possessions" in place of "personal information" [111, 211, 288]. This shift signals an expanded focus that includes new types of data:

social network data, location data, lifelogging data, metadata, and so on. With new devices such as voice and virtual assistants or IoT (Internet of Things) products coming onto the market, even more types of data will become part of people's life and feel increasingly personal.

In this dissertation, we use the term *personal data management* to take a broad focus that includes but is not limited to FM and PIM (Figure 2.1): personal data *is* information, but it is not only files or folders, and not all data is information. For example, data can include articles from online magazines, as one participant in Study 1 mentioned (these are not strictly personal since they are public), or mobile applications (better known as "apps," these are perceived as different from files and folders), or computer logs (this are not necessarily perceived as information).

Our work largely focuses on *digital* data. In Study 2 (Chapter 5) we touch on key differences between physical and digital items that can inform the design of tools for digital data curation.

Throughout the dissertation, we touch on data captured automatically by technology. However, we do not delve into the background of lifelogging technologies (i.e., Personal Informatics systems), which we see as a related but separate area of work within HCI that previous studies have explored in depth [85, 87, 148, 163, 308].

2.1.2 The moral economy of personal data management

In looking at personal data management, we use two theoretical frameworks to define the scope of our our work. The first comes from Vertesi *et al.* [288], who detail how personal data can take many forms (from photos to files and logs), living in *ecosystems* made up of multiple devices and relationships, with online cloud storage platforms becoming increasingly prominent. When managing their personal data, people face a tension between sharing *with* and safeguarding data *from* others. People often make decisions based on moral convictions about what they think is the "right way" to manage data. We adopt this *moral economy of data management* as a backdrop for much of the analysis in the dissertation.

In our work we touch on privacy issues, but do not focus on sharing behaviors. Sharing data with other people is without doubt an essential component of modern user practices in this space and undoubtedly affects some curation decisions. However, to scope the focus of our investigation we do not delve into collaborative data management or sharing practices. Studying this related but different aspect of data management would require different methods and samples. We refer to work by Massey *et al.* [186], Voida *et al.* [293], and Rader [239] for more details on collaborative user behaviors.

2.1.3 The personal data curation cycle

The second framework helps us narrow the focus of our work. We refer to Whittaker [302] to reframe common PIM actions as part of an "information curation" cycle consisting of three stages (Figure 2.2): *keeping* (where users acquire or create information and then decide what to keep or discard), *managing* (where users organize the items they keep using folders or other structures), and *exploiting* (where users retrieve items by searching for them or navigating in their structures). (In combining the two frameworks, we replace Whittaker's use of the term "information" with "data," as explained above.)

We largely focus on the first stage of curation, where users acquire or create data and then decide what to keep or discard. Throughout the studies we also touch on the second stage, management, looking at how keeping decisions intersect with organization. Our focus on the last stage of curation, exploiting, is limited. Information retrieval represents a sub-field of its own, with a large body of work covering strategies for re-finding data and providing new tools for improving user practices. However, the retrieval stage still informs some of the ideas that we discuss in several studies. Where relevant, we point out specific past work and insights about data retrieval.

The terminology of personal data curation

One challenging aspect of studying data curation is the lack of standard definitions for overlapping ideas and user actions. Archiving, curating, preserving, managing, organizing, storing, cleaning up, selecting, decluttering: these are all terms that appear in previous work. But often, researchers use them inconsistently and in a way that might differ from users' interpretations (many participants we talked to,

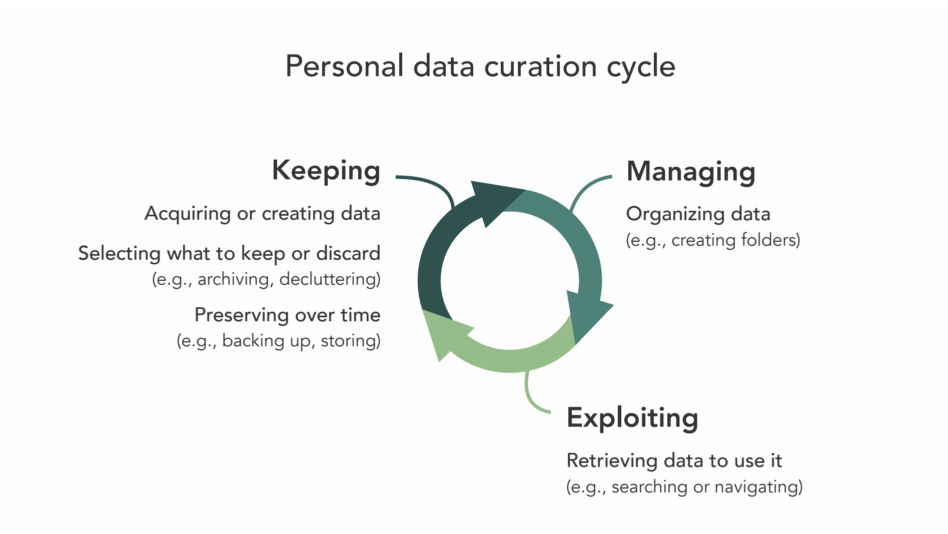


Figure 2.2: An illustrative elaboration of Whittaker's personal data (or information) curation cycle, consisting of three stages [302]: keeping, managing, exploiting. In our work, we focus on the first stage, keeping, elaborating on possible actions that people take. This illustration lists some of the terms and actions that we focus on.

for example, used the word "archiving" to indicate they wanted to hide something, whereas research on "digital archiving" often uses the term to indicate intentional long-term storage). The messy nature of the language of personal data and its consequences for user behaviors is a discussion item that we bring up later in our work, proposing directions to address it (Chapter 7 and Chapter 8). On our part, as we note in the Preface to the dissertation, we use different terms to highlight specific aspects of the curation process that overlap but can be framed as slightly different. Even though this approach might contribute to the confusion in terminology, we think it is important to capture the nuance of different actions in personal data curation.

Throughout the dissertation, we use the expression *keeping and discarding decisions* to indicate the first stage of curation, which is the key focus of our work. Within this stage, we elaborate on possible actions that users might take after acquiring or creating data. (Figure 2.2) We name *data selection* the overarching process of selecting specific items to keep or discard. After making this selection, users might engage in *data preservation* for items they want to keep (i.e., taking additional actions to preserve items in the long-term, such as doing backups or storing items in specific places) or *data decluttering* for items they do not want to keep (i.e., getting rid of items or re-organizing them–this is an action where two stages of curation, keeping and managing, are closely linked and often happen simultaneously because by re-organizing it becomes more clear what to get rid of). We see *archiving* as an overarching and somewhat ambiguous action that can be used both to preserve and declutter items depending on what meaning the action has for each person. The concept of a *personal archive*, instead, is the general idea of a set of items stored somewhere. Engaging in personal data curation should help make that personal archive feel more meaningful and intentional.

We focus on *preservation tendencies* in Chapter 4, *criteria for decluttering* in Chapter 5, *support for selection* in Chapter 7, and *curation as a whole* (but still with a stronger focus on keeping and discarding decisions) in Chapter 6 and Chapter 8.

2.2 Practices and challenges in personal data curation

Previous studies report how people engage in personal data curation for several reasons: for example, to build a legacy [111, 112, 149], support memory [140, 153], manage and honor relationships [153, 288], find things later [149], or present their image to others [211, 315]. In some cases people actively acquire or create data [24], sometimes as part of an ongoing collection [89, 299]. In other cases, they accumulate data more passively [24, 307].

2.2.1 Keeping data to build an identity and to remember

One key reason for keeping and wanting to preserve data is to build an identity [66, 68, 149, 211]. As with physical possessions [17, 65], people can form attachments to digital items [18]—a phenomenon called "self-extension" [17, 18, 66–68]. A key way for possessions, whether digital or physical, to help shape identity is by supporting memory and reminiscence [149, 153, 229]. Sentimental possessions (e.g., photos in particular [43, 44, 152, 218, 307]) can help people remember key moments and relationships from their past, acting as symbols of key experiences in their life [153, 228]. Much of previous work explores the idea of digital "mementos" [39, 147, 230, 278] or "heirlooms," [10, 210, 212] investigating ways to incorporate digital data and its potential for reminiscence into everyday life and into people's homes [217, 218, 226].

However, previous research shows how not all "stuff" is created equal. In the case of digital data, only some items feel as possessions [68]. To determine the importance of digital items people might refer to values such as utility and recency, emotional attachment, craft, and replaceability [68, 104, 153, 210]. Sometimes these values are shared between physical and digital items, but people can perceive digital items as less unique because they lack the material qualities of physical items [104, 210].

The literature on material culture suggests that disposing of possessions is as important a practice as acquiring them [60, 122, 159, 196, 249]. Studies on physical "decluttering stages" [59] or "selection regimes" [140] report useful accounts of selection practices, showing how sorting and discarding physical items can feel enjoyable compared to sorting through digital items [227].

But when it comes to personal data, many users find the act of deleting items "onerous," a "burden," [140] to the point of questioning the idea of deleting digital items in the first place because it is against "their nature" [111].

2.2.2 The difficulty of discarding data

The ongoing process of curating data is challenging because it often involves anticipating the future [24, 302]. Many users end up choosing to keep all their personal data as a default [38]. Bergman & Whittaker [24, 302] refer to *prospect theory* [145] to explain why. In general, people tend to be averse to risk and perceive a potential loss from discarding data as more substantial than any gain. (i.e., the possibility of losing information that turns out to be useful outweighs the convenience of finding data more easily.)

There is also evidence that digital tools offer poor support for curating data besides going through items one by one [130] and that users can have confused mental models of how deleting works, especially in cloud platforms [240]. The result of these challenges and perceptions is that many technology users "approach long-term preservation [of data] with trepidation" [227].

The prevalence of keeping as a default [38, 151], however, creates an accumulation cycle: the more data there is, the harder it is to select, organize, or find things [24, 307]. An emerging body of research points to people feeling stressed by the accumulation of data [253, 271].

The increasing popularity of cloud platforms further complicates keeping decisions. People can perceive their digital data as being an undefined collection of items without knowing exactly what they possess and where things are. As Odom *et al.* [213] remark, "the role of curator can become complicated if one does not know what one is curating." Yet, there is a desire for some form of selection. Previous studies highlight that a meaningful personal archive should focus on "the remarkable" items [166] and that selection is one of the ongoing practices that inform what an archive is [149]. Work on the gaps between actual and desired PIM behaviors highlights how users are "eager" to delete unnecessary information and would like to be more efficient in this practice but often fail to do so because of poor support in data management tools [5].

2.2.3 Previous insights about discarding decisions

Previous HCI studies mention several reasons for deleting or discarding data. For example, users might want to delete data from an ex-partner after a breakup [247], unneeded, unknown or unwanted data [150, 184, 213], data copied online [211], social media accounts or expired accounts [288]. Most of these studies, however, mention these episodes as secondary insights. Our work looks more closely at discarding practices in the context of personal data curation, providing a richer characterization of user behaviors.

Some of the more detailed studies about deleting practices look at the problem from a psychology or security perspective. For example Kim [151] reports that people might decide to delete for a variety of reasons and with a variety of approaches: because of limited space on their devices, after buying a new device, by inspecting items one by one, accidentally, or more regularly. Kahn *et al.* [150], instead, investigate retrospective data management decisions in two prominent cloud storage platforms, Google Drive and Dropbox. They show that most participants would like to delete at least one from a set of 10 old files when given the opportunity to review them. Ramokapane *et al.* [240] also look at cloud platforms and list the main reasons for wanting to delete data: among them, privacy protection, storage issues, external company policies, and the perceived value of the data. They also mention how some users decide to delete for the sake of it, to "tidy things up."

Additional studies look in detail at deleting decisions for emails [202], data on social networks [4, 202, 260, 298], and messages [251]. With email, reasons for deleting tend to focus on storage capacity and information utility, with users reporting deleting emails that are old, unneeded, or taking up space [202]. With social networks, instead, privacy and regret play an important role, with users wanting to remove embarrassing content [4, 202, 260, 298]. A key reason for deleting messages, instead, is to fix and revise exchanges in a conversation (e.g., typos, inappropriate content, obsolete content) or free up space taken up by media content [251]. But emails, social networks, and messages add complexity to deleting decisions because other people are involved: even if a user decides to delete a message, for example, the person they are chatting with might have seen it already [251].

Past work also makes a key point: user interfaces should avoid "one-size-fits-all" solutions for supporting deletion because deleting decisions are contextual and hard to generalize [150, 202].

2.3 Summary and conclusion

Overall, previous work shows that personal data curation is an important activity for people, but the literature lacks a rich and systematic understanding of individual decisions for keeping and discarding data. Previous work pays less attention to individual differences or design efforts in the initial stage of curation, as compared to organization and retrieval. Deciding what to keep or discard is a challenging aspect of curation where users often struggle because of the personal and contextual nature of these decisions, but also because users lack proper support in the tools they use to manage data.

Our work adds to the existing literature by shifting the focus from data organization and retrieval to decisions about what data to keep or discard. We advance existing literature in several ways. First, we detail individual user preferences for discarding decisions in different contexts (Chapter 4, Chapter 6), and categorize criteria and strategies for discarding data (Chapter 5). Then, we explore how to move beyond a one-size-fits-all approach and incorporate individual preferences both in contextual but separate design solutions (Chapter 7) and with a cohesive, personalized tool (Chapter 8).

In the following chapters, we will touch on additional related work and highlight differences between our work and previous studies.

Chapter 3

Methodology

In this chapter, I discuss my research approach and methodology. I outline criteria for evaluating my work and reflect on my research positionality. *Reflexivity* is an important component of rigorous qualitative research: it informs the reader about possible biases or influences on the analysis, and it positions the researcher as an active actor in the research process [56]. Because reflexivity emphasizes the role of the researcher, I write this chapter in the first person.

3.1 Research approach

My research approach is broadly speaking *qualitative*. Over the course of my work I combine two philosophical research paradigms: *constructivism/interpretivism* and *pragmatism* [197]. Combining approaches and methodologies is common in HCI, a multi-disciplinary field that embraces many perspectives.

Study 1 and 2 fall largely within a constructivist and interpretive paradigm, where the goal is to understand and interpret a culture-bound reality [197]. Across the two studies, I use a constructivist grounded theory methodological approach [56]. Grounded theory is a methodology for collecting and analyzing data in a "systematic but flexible" way [56]. Grounded theory originates in social science, but it is a common approach in the field of HCI [201] (although HCI studies often only follow a light version of grounded theory, without using the full set of procedures). The goal of grounded theory is to build an understanding of processes and phenomena grounded in research data. Key steps in grounded theory include: *theoretical sampling*, for selecting participants based on gaps in an emerging theory rather than to build a representative sample; an initial phase of open coding (often going through interview transcripts line-by-line), followed by axial and selective (or focused) coding (where the researcher focuses on a few key ideas); a process of *constant comparison*, where examples and experiences from different participants are compared to help better understand the process at hand and its key categories; *memo writing*, where the researcher engages in reflections about the analysis. There are different schools of grounded theory, a result of years of practice and debate among scholars, with diverging views on the nature of qualitative research [54, 55]: at a high level, we can distinguish between *objectivist* grounded theory (as conceptualized by Corbin & Strauss [267]) and *constructivist* grounded theory (as conceptualized by Charmaz [56] and, before that, in the original conception of grounded theory by Glaser & Strauss in the 1960s [103]). Charmaz summarizes the difference between objectivist and constructivist grounded theory explaining that "constructivist grounded theory assumes relativity, acknowledges standpoints, and advocates reflexivity" [55]. In my work, I follow Charmaz's school of grounded theory.

My work moves closer to pragmatism with Study 3 and 4, where I take a *Research through Design* (RtD) approach [317]. At a general level, RtD argues for integrating design into research, giving emphasis to creating design artifacts that can drive research inquiries. However, the broad nature of RtD often includes diverging methods and perspectives [96, 318]. In my work (especially in Chapter 7), I borrow more closely from methods such as *Speed Dating* [72], *User Enactments* [214], and *Experience Prototyping* [45].

Speed Dating is a method that argues for exploring a variety of design ideas and dimensions, without requiring full technical implementation [72]. User Enactments, instead, ground the evaluation of possible design ideas into situations and contexts from a potential future, allowing participants to experience alternative scenarios [214]. Finally, Experience Prototyping argues for creating prototypes that can simulate user engagement with a product, often in a set of scenarios that might be difficult to experience directly otherwise [45]. In my design work (Chapter 7, Chapter 8), I combine aspects from these methods and adapt them to the specific requirements of my research inquiry.

In terms of data analysis, I refer to *reflexive thematic analysis* (TA) [61] in most of my studies. Reflexive thematic analysis differs from other schools of thematic analysis ("codebook" thematic analysis and "coding reliability" thematic analysis) because it gives emphasis to the researcher subjectivity and sees coding data as a reflexive, recursive process [41]. As Braun and Clarke explain,

> "Quality [in] reflexive TA is not about following procedures 'correctly' (or about 'accurate' and 'reliable' coding, or achieving consensus between coders), but about the researcher's reflective and thoughtful engagement with their data and their reflexive and thoughtful engagement with the analytic process." [41]

Thematic analysis, however, is a *method* and not a *methodology*: this is why I use it both when I refer to a grounded theory approach and also when I refer to a research through design approach.

In general, my work moves away from post-positivism and experimental research, where common goals are predictability, statistical generalizability, and representativeness. Instead, my work is exploratory, holistic, generative, and descriptive:

- *Exploratory* because I openly and inductively explore a specific topic that the existing literature has not covered in detail (i.e., people's decisions about what personal data to keep or discard). There are no *a priori* hypotheses to test here [175, 266].

- *Holistic* because I take a broad perspective on the topic, letting participants guide me towards many different types of data, devices, platforms, and tools that they consider part of their everyday life. There are no imposed definitions of what should be part of the research. Instead, the focus of my work is grounded in participants' daily experiences.

- *Generative* because I use the varied sample of participants and the resulting insights for generating design ideas and open a space for new opportunities.

My goal is not to provide a statistically representative account of behaviors that can automatically generalize to a broader population, but rather my analysis is grounded in a specific context that informs its *transferability* (see below).

- *Descriptive* because I strive to provide a rich, thick, contextual description of people's behaviors but I do not claim the ability to precisely predict future behaviors based on my insights [110].

Why do I take this approach? It matches my world view. I believe that methodology should follow from *ontology* and *epistemology*. These are key philosophical branches that tell us *what* we can know about and *how* we can know about it [63, 197]. They inform research according to a personal stance. My stance is that the phenomena I study should be approached with a perspective closer to social science, rather than natural science [197]. Thus, in my ontology I refer to *bounded relativism* (i.e., constructions of reality are bound to cultural and social contexts), and in my epistemology I refer to *constructionism* (i.e., people *construct* reality). The resulting philosophical perspective I take is *interpretivism* [63, 197].

My stance does not imply that a quantitative approach would be wrong and a qualitative approach is the only possibility for studying the phenomena I look at. Simply, a qualitative approach is what matches *my* philosophical perspective on the topic. Another researcher could look at the same topic with a different philosophical perspective and then take a different approach.

A qualitative approach also stems from to the nature of the research questions I am interested in exploring with my work. The research questions I highlight throughout the different studies focus on new and uncharted research territory. Thus, a qualitative approach is the appropriate methodological fit for my work, where the goal is to develop a rich, descriptive understanding of emerging areas of interest. A qualitative approach helps in revealing and understanding salient issues for future research. Rich, one-on-one user interviews help understand the current state of user behaviors, opening opportunities and directions for more focused work. Research through Design methods, instead, help in probing future states, anticipating how design can change and influence current situations.

3.2 Criteria for quality

Criteria for this type of qualitative work include rich rigor, credibility, resonance, and sincerity [280].

- *Rich rigor* considers the complexity of the work, its face validity, and carefully chosen procedures for data collection and analysis [280].

- *Credibility* considers the "trustworthiness, verisimilitude, and plausibility of the research findings." [280]

- *Resonance* considers the aesthetic value of the text and its *transferability* [110, 165] (i.e., the possibility of transfering results to a similar but separate context).

- *Sincerity* refers to self-reflexivity and transparency in the research process [280] (as outlined below).

3.3 Research positionality, reflexivity, and transparency

A common practice of qualitative research, and constructivist grounded theory in particular, is to outline the researcher positionality, providing a reflexive account of key decisions in the research process. An additional ethical requirement in many academic journals is to acknowledge a potential conflict of interest that influences the work reported. These statements are unfortunately rare or non-existent in HCI publications. I include them to lend rigor and transparency to my dissertation.

3.3.1 Research positionality

When the subject of my dissertation comes up, people often assume that I am a hardcore minimalist on a mission to purge everyone's devices. Hide your hard drives, save your data. But no. I do spend time regularly managing and curating my data, but like many of the participants I talked to, my approach is contextual and varied. For example, on my phone I have collected close to four years of mood tracking and money tracking data. I will probably not delete this information any time soon and I do regular backups, a practice I started after losing all my data more than a decade ago. On Spotify I have a collection of playlists that I have been putting together month by month for over three years. On the other hand, I strive for a clean email inbox, I structure the organization of my documents, and I try to keep the storage space I use on my computer and cloud platforms to a minimum. I grew up before paying for digital things became an everyday practice and to me digital still means free. I do not want to pay for storing my data. These experiences, practices, and attitudes inform my research and the specific perspective I take in looking at personal data curation. At a general level, my position informs the idea that studying data curation is important, as discussed later on in Chapter 4. It also helps in noticing and observing curation behaviors that differ from my own. Because this is a topic I am fascinated by and I have first-hand experience with, I am alert to individual differences that might go unnoticed or appear as meaningless to other researchers.

3.3.2 Reflection on my analysis choices

The qualitative approaches I refer to in my work place great importance on the researcher as the key instrument of investigation. Reflecting on the decisions that drive the research process is key for understanding the resulting analysis. Here, I highlight some of the most important decisions that determined the outcome of my research.

I did not plan my PhD to be about personal data curation. In fact, the original research question of Study 1 was to investigate how users approach backups. I wanted to know whether people still do backups and how. But then, during the study, I found the idea of "hoarding" and "minimalism" as two opposite tendencies more interesting. I decided to follow this lead and shift the focus of the study, influencing everything that came after. I could have recorded the insights about hoarding and minimalism along others, give them no particular emphasis, and discuss other themes based on the data collected. But I made a choice that my position and perspective as a researcher informed.

Similarly, in Study 4, I could have detailed how participants used each section of the prototype we evaluated, how they ranked each and every tool we asked them about. But I did not think those were necessarily the most interesting insights, so when I identified the idea of data boundaries from one specific participant quote, I decided to follow that conceptual lead and structure the analysis around this one key idea. Another researcher might have approached the same data differently. Interviews are rich by nature and capture a lot of information; this is one of the advantages of this method. But at some point, as a researcher, you have to decide the focus of the analysis–decide what story to tell. In this dissertation, you will ready the story I have chosen to tell.

Throughout my PhD work (starting from Study 2) I regularly engaged in journal and memo writing, a common practice in grounded theory [56] that helped me reflect on my analysis as I went along. Several ideas I wrote about in my memos eventually made it into the analysis, but others did not. Appendix C includes some examples of memos that might be helpful to gain more insights into how my thinking and process around specific ideas evolved.

Appendix C also includes some coding examples from multiple studies. I provide them for transparency but with an attached disclaimer: a set of codes does not make an analysis. It is tempting to place too much importance on them, and transfer any responsibility in the analysis to "the codes," maybe going at lengths to produce a codebook that can be routinely applied to the data. This is not a view of qualitative research I share. As Charmaz argues [56, 101], codes are a tool to understand and stay close to the data, but at some point the researcher needs to move from codes to a more abstract and conceptual understanding of the data. Concepts and themes are not necessarily captured in a single word or a neat correspondence between a paragraph and a code: that is where the actual analysis comes in. The more interesting themes and concepts often cut across codes, categories, participants. Like Braun and Clarke [41], I do not believe that themes and concepts magically "emerge" from data, as if they were there all along, hidden, waiting to be found. Qualitative analysis requires interpretation. Interpretation requires choices about what to focus on and what to ignore. Coding is the first step in this process but it is not an end in itself.

3.3.3 Conflict of interest disclosure

Finally, I see it as an ethical obligation to acknowledge my internship as a User Experience Researcher at Google in 2019. While my work at Google was entirely independent from my PhD research, Google and its products often come up in my dissertation because of their relevance to the topic of personal data. Throughout my work I keep the position of an independent researcher and keep my experience at Google separate, expressing opinions on related products based only on my work and its implications. However, I am aware that bias is often implicit and having worked at Google is a potential conflict of interest in the context of my PhD work. I hope that this disclosure can make my position more transparent and help readers fairly interpret my work.

Chapter 4

Study 1

Identifying a Spectrum of Data Preservation Tendencies

In this first interview study[1], we focused on exploring general tendencies around what data participants decided to keep or discard over the years. We used the term *data preservation* to indicate the practice of keeping data in the long-term, and we see this as a key component of the *keeping* stage in the personal data curation cycle. This first study details the nuance of people's practices in this space and the importance of curating data for identity construction: this key idea will come back throughout the dissertation and represents the core underlying theme of my work.

4.1 Introduction

Economists argue that digital data has become the most valuable resource of the 21st century [276]. Like oil, it is a resource that big companies are trying to control and extract from people in large quantities, because it drives economic transactions [264]. Every day people produce, store, share, and interact with an increas-

[1]Originally published in Vitale, Janzen, and McGrenere. (2018) *Hoarding and Minimalism: Tendencies in Digital Data Preservation*. Proceedings of the 2018 SIGCHI Conference on Human Factors in Computing Systems (CHI '18) [290].

ingly large amount of data, including pictures, texts, files, mobile applications and the data they contain.

Cloud platforms are one of the solutions that leading technology companies have proposed to deal with the increasing amount of data. These platforms often cause confusion [213] and raise privacy concerns [138, 288], but they offer seemingly unlimited storage that requires little maintenance on the user side. This explains why they are an increasingly popular choice to store digital data for everyday users [288]. Storage is either cheap or outright free. Google Photos, for example, offers unlimited space for pictures (although at reduced quality).

This is the "seductive" digital landscape that Marshall [182] predicted a decade ago when studying long-term preservation of digital items. At the time, a similar change was taking place: hard drive storage was becoming cheaper, giving users the option to store nearly "everything" [184]. The pervasiveness of the cloud is once again reinforcing this possibility. Now that people are living in this seductive landscape, how are data preservation practices changing? It is critical to understand how users are experiencing this new world, as they are just in its foothills. As storage gets cheaper and digital data more of a commodity, how do users deal with this new environment?

In this first chapter, we are interested in the act of *preserving* (or selecting) data, by which we mean deciding what data to keep and discard. As Whittaker [302] points out: little is known about "when and why people keep or delete different types of information." Therefore, we focused on a main, broad research question: how do people approach digital data preservation in the cloud age? How do they decide what to keep and discard?

We interviewed 23 participants from diverse backgrounds, focusing on their current and past digital data practices. We asked them what "stuff" they kept through the years and why, how they used it, what they considered important, and how they made sure not to lose it.

We found that participants approached data preservation driven by a range of underlying tendencies, living on a spectrum with two recognizable extremes: hoarding (where participants tended to accumulate a lot of data even if it had little value, rarely deleting it) and minimalism (where they avoided storing too much data or regularly engaged in a cleanup process).

First, we characterize in depth the spectrum and its extremes, showing the nuanced nature of preservation tendencies. Then, we compare and contrast different preservation strategies, focusing on the extremes of the spectrum, elaborating on how they helped participants build their identity, a practice commonly associated with possessing data [149]. Finally we discuss, among other things, how our categorization relates to previously reported behaviors (e.g., filing and piling [177], email cleaners and keepers [117]) and the broad implications to shape the current technological landscape.

4.2 Related work

Here we report related work on data preservation and digital hoarding, expanding on some the studies cited in the Chapter 2.

4.2.1 Framing data preservation

We use the expression *data preservation* to indicate a subset of what is commonly thought of as data management or curation, as others have done before [156, 182]. In the overall process of data curation, preservation looks specifically at deciding what data people might choose to keep.

Although preservation is often overlooked in favor of other curation stages [302], there are clues in previous literature about general practices and values users refer to. However, we argue that there are gaps in the current literature.

Previous studies show that users tend to take a neglectful approach to preservation: they do not think carefully about long-term preservation expecting data to somehow survive without planning [184], they have inconsistent strategies for short-term preservation with a mix of "planned" (e.g., doing a manual backup onto an external hard drive) and "unplanned" methods (e.g., emailing documents to other people as part of other activities) [156], they make no clear distinction between short-term and long-term preservation, using terms such as "storing," "archiving" and "backing up" interchangeably [156, 184, 263].

When people preserve data, they do so, among other reasons, to build an identity [66, 68, 149, 211], as we have discussed in Chapter 2. In this chapter, we extend the idea of building an identity and further explore data values that people might refer to[2].

What is missing in the current literature on data preservation is a holistic understanding that can explain the broader context of these values against changing technologies. While insightful, previous studies often either focus only on computers [156] or specific populations (e.g., academics [149], photographers [263]), or predate the current technological landscape and its significant changes [149, 182, 184]. With a broader approach, we aim for a more comprehensive understanding of user practices.

4.2.2 Digital hoarding

Digital hoarding is not an entirely new phenomenon, but little is known about it. Coming from a background in psychiatry and neuroscience, van Bennekom *et al.* [284] present a clinical case of digital hoarding with one patient who suffers from a hoarding disorder that leads him to take 1,000 pictures every day. They define digital hoarding as the "accumulation of digital files to the point of loss of perspective, which eventually results in stress and disorganization." They also propose to categorize digital hoarding as a sub-type of the hoarding disorder and point to the lack of scientific papers on the subject. The topic is just now gaining interest in the broader scientific research community, as evidenced by additional studies published starting in 2018 [205, 206, 222, 253, 271].

In a review of published literature, Gormley and Gormley [105] discuss in general terms the costs associated with data hoarding and digital clutter based on previously published literature: for example, hoarding data can result in costs for storage space and management overhead. However, the research literature on the subject is extremely scarce compared to hoarding of physical objects, a much more widely studied phenomenon, with tools to measure it and diagnose it [95]. In addition, all of these studies are from outside the HCI literature, where these terminologies are not well recognized and only mentioned in a few studies [126, 127, 250]. Before running our study and encountering hoarding and minimalism, we were not aware of research on the subject. We note, however, that we refer to hoarding as a set

[2] In the following studies, and particularly in Chapter 5, we expand on criteria for deciding what data to discard.

of everyday tendencies, not as a disorder. We are not in a position to diagnose participants.

4.3 Methodology

4.3.1 Participants

We interviewed 23 participants (16 females, 7 males) in Vancouver, Canada. We used purposeful sampling to gather a relatively varied sample in terms of age, ethnicity, and background. Participants' ages ranged from 21 to 64 (average: 35.4, median: 30, SD: 12.5). Occupations included business consultant, cook, mental health worker, server, researcher, research coordinator, retired accountant, special educator, social worker, software tester, stay-at-home parent, trader, university-level coordinator, in-between-jobs, in-between study and work, full-time graduate and undergraduate students (8 with backgrounds in Architecture, Archiving, Commerce, Education, Electrical Engineering, Kinesiology, Mechanical Engineering, Organizational Behavior), part-time graduate students who were also working (2 with backgrounds in Arts and Gender studies). The majority of participants (15) had basic technical skills, followed by average (4) and above average (4). Participants were compensated $15 each.

4.3.2 Procedure

After conducting three pilot interviews, we recruited participants through mailing lists and posters in several community centres in the city. We conducted semi-structured interviews, each lasting on average 45 minutes, at a location chosen by participants. One member of the research team conducted all interviews. We asked participants to bring their main interactive devices (e.g., laptop, smartphone, tablet) to the interview. All interviews were conducted in English. We recorded the audio of the interview, took hand-written notes, and later transcribed them for analysis.

4.3.3 Data collection

After collecting demographic information, we asked participants to talk about and show us their digital data, whether it was files, data from mobile applications, or other examples. Following the example of Vertesi *et al.* [288], we did not impose a specific definition of digital data. However, unlike Vertesi *et al.* [288] we asked participants to *show us* the data, although they were free to choose what to show so as to respect their privacy. Participants gave an overview of their devices and then a more detailed tour of each, explaining what they used them for, what data they had on them, and why they had kept it. Then, we asked participants to imagine what they would want back if their devices broke down or were stolen, focusing on what was the most important data they had on them and why they considered it important.

The second part of the interview revolved around a light version of the life history method, a technique used to relive the life an individual through their narration [181]. In the context of our study, we asked participants to relive their digital life history, focusing on the devices they used through the years, asking them to remember what they used them for, what data they had on them, whether they kept it or not, and why. We also encouraged them to sketch a chronology of their digital life history on paper to help them think through it (Figure 4.1). Life histories are useful to understand individual experiences in the social context where they take place and how individual understanding evolves through time [181]. During the interviews, we used the sketches as prompts to encourage participants in discussing how their data evolved over time. Then, during the analysis process, we referred back to the sketches to contextualize participants' recollections.

We also asked participants to think about their data one and ten years into the future, to know what they anticipated as something worth keeping and why. We concluded by focusing on positive and negative aspects of their data management.

4.3.4 Data analysis

We analyzed the interviews using the Braun and Clarke approach to thematic analysis [61], where "coding is flexible and organic and evolves throughout the coding process" [62]. We did both an inductive and deductive analysis (based on the "data economy" framework [288]). We used open coding with all members of the research team examining and discussing the data collaboratively in an iterative and reflective process. Each member could see how others were coding the data,

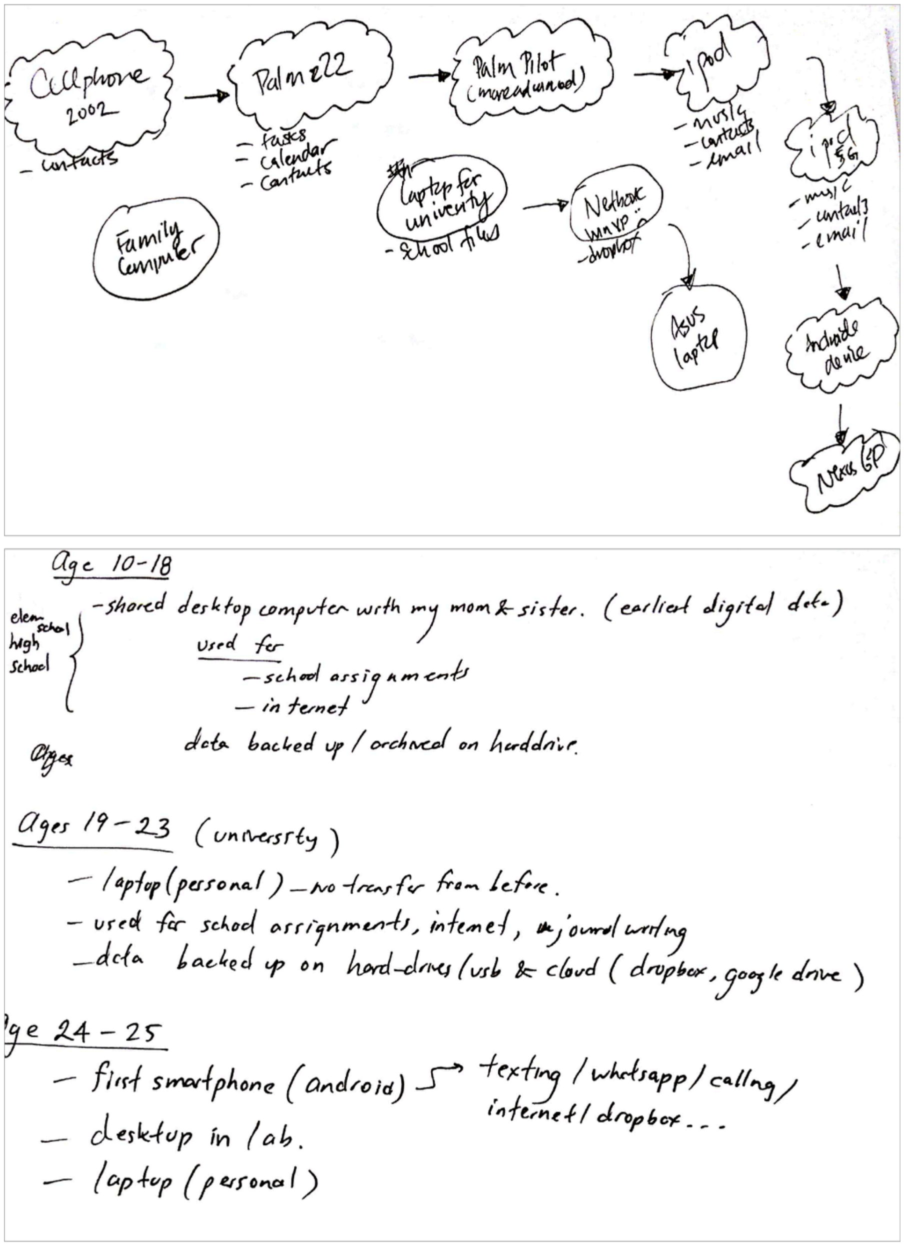


Figure 4.1: We asked participants to sketch the history of their data over the years to see what they had kept or discarded. In this figure, examples of sketches from two participants.

discuss the interpretations, and propose alternative explanations. Throughout this process, we regularly met for lengthy in-person discussions of the interpretations, making sure they were coherent, comprehensive, reasonable, and reflective of the actual data. We later grouped codes into categories and went back again to the interviews to check for consistency.

We looked at preliminary trends after the first batch of interviews, adjusted our research *foci* and proceeded with additional interviews until thematic saturation. That is, when we got to P20, we noticed interviews were starting to closely repeat ideas from previous participants, therefore we stopped at P23. In the later stages of the analysis, we paid attention to the contrast between hoarding and minimalism. These terms came up halfway through the study, when some participants used them to describe their approach to data preservation.

We do not present counts for specific occurrences of behaviors, as we focus on recurring patterns of behaviors *across* and *within* participants to characterize hoarding and minimalism. We agree with Braun and Clarke that "frequency does not determine value" [62]. Our goal in reporting is to characterize the essence of hoarding and minimalism, not their distribution, as our methods simply do not allow us to give a distribution.

Epistemological stance and reflexivity

In our analysis, we took a constructivist epistemological stance within a bounded relativist ontology [197]. In the context of HCI, we position ourselves in the so-called third paradigm, where meaning-making is a central focus [124].

Our approach is similar to a constructivist grounded theory approach [56]. Taking a constructivist approach means that we saw interviews as an interactive process of meaning-making: we built knowledge with participants. Therefore, we do not claim absolute truths about people's behaviours, but a shared understanding grounded in participants' reasoning and experience, reflective of their broader cultural environment. Focusing on the words used by participants, we arrived at the notion of hoarding and minimalism. These terms are socially constructed in the sense that they embody cultural connotations: we debated whether they were appropriate, reflecting on our assumptions about what they point to. Ultimately, we use them to fairly represent the shared understanding we constructed with participants.

In line with our constructivist approach, we critically reflect on our position as researchers and its influence on the analysis. Throughout the analysis, we reflected and discussed our own experience with data preservation, since it is something we deal with on a regular basis. In particular, one team member considered themselves to have mostly minimalist tendencies, one had a mix of both, and another one reflected on a tendency to hoard pictures. Additionally, we frame data preservation as a challenging task worth investigating but others might have different perceptions. We also acknowledge that our Western cultural background and its values inform our view. This points to the inherently interpretive nature of our work taking place in the the current socio-technical landscape.

4.4 Results

First, we present contextualizing information about the data that participants discussed in the interviews. Then, we focus on the cross-cutting theme of hoarding and minimalism, giving an overview of recurring behaviors across participants. [3]

4.4.1 Contextualizing information

Similar to what Vertesi *et al.* [288] found, participants considered a variety of data sources: computers, smartphones, tablets, wearable devices, online platforms, and mobile applications. They talked about files, text conversations, pictures, videos, bookmarks, logs, profile settings. Pictures were consistently regarded as one of the most important pieces of data because of their sentimental value. Participants mentioned how photos served as a tool for remembering and how it would be hard to take them again if they lost them: "*I can't retake those photos* [...] *I'm emotionally attached to them* [...] *Music, I can always download again. It seems like photos are less replaceable.*" (P3) Other factors determining the importance of data were recency, utility, time invested to craft it, and its role as a record. We are not the first

[3]The two other themes were "the importance of other people" and "striving for the right way to manage data." We decided not to include them in the final analysis because they had substantial overlaps with work by Vertesi *et al.* [288]. In Chapter 6 we touch on the importance of other people and how it helps in understanding data curation.

to report these values [104] but we elaborate later on the important role of data as a memento in relation to hoarding.

4.4.2 Identifying hoarding and minimalism

Halfway through the data collection process we met with Sarah (pseudonym for P13), the participant who introduced us to the approach of data minimalism.

Sarah is a graduate student and mental health worker. She manages all her data on her laptop. She does not use cloud platforms. She has a phone, but it is not a smartphone. It is a "dumb" flip phone. She explains that she grew up in a small community whose members did not use tablets or smartphones. She does not want one, "*never*", because "*otherwise* [she'll] *be on the bus* [demonstrates hunching over the phone]" and instead she wants to "*look outside and talk to people.*"

On her laptop, a small MacBook Air, there is only one main folder simply called "Life." "*Everything is kinda organized,*" she explains. "*My apartment, inspiration, beautiful photos, photos of my family.*" When we ask her why she called it "Life" she takes a moment to think. "*Well, I was thinking about it.* [...] *And I was like, OK, what could this be? Well, it is my life. My family, my school. I mean, my life is so much more than that. But I couldn't think of a better name.*" She then explains how data helps her build an image of herself, an idea that will become important to understand the broader role of preservation tendencies:

> "I think humans are always trying to find things outside of themselves to make them feel they're more than they are. If I like a song, it's part of me, me kind of building up the image of myself. So I think it's me being like 'Oh yeah, my life, school and this and that.' We don't need things outside ourselves, but we are always looking for things to make us feel complete." (P13)

At the end of the interview she summarizes her approach to deciding what to keep and discard: "*It's very minimal. I try to delete everything that I don't need as fast as I can.* [...] *Do most people have a lot of stuff?*"

Yes, they did. Compared to what we had observed up to that point in the study, hers was a very different approach. In retrospect, it was clear that until that point we had mostly seen strategies closer to another extreme: hoarding. We thought Sarah might be a "unicorn" and that we would not meet other participants like her. But we did. And then something similar happened when other participants self-identified as hoarders, even though we never used these terms in our questions. In fact, looking back, we saw that some participants had specifically mentioned "hoarding" before we interviewed Sarah, but it had not jumped out to us. Altogether, it became more and more clear that participants adopted a range of data preservation tendencies that lived across a spectrum between hoarding and minimalism.

We start by describing and characterizing the tendencies participants reported, largely grouping them along the two extremes of a spectrum: hoarding and minimalism. Throughout the analysis, we point to instances of nuance within individuals, with some participants being highlighted in both sections on hoarding and minimalism, or displaying interesting exceptions to their general approach. Broadly speaking, some participants were stronger in their tendency towards hoarding or minimalism, while others displayed a much more even mix of both or were not easily classifiable. However, this is an over-simplistic categorization, given the nuanced nature of the tendencies and the fact that they represent a spectrum of behaviours within two recognizable extremes.

We also touch on the actual organization of data that participants displayed (e.g., being organized or messy, using folder hierarchies or not), showing that it appeared to be orthogonal to their tendencies—some participants were organized, some were not, independent of the tendencies they displayed.

4.4.3 Hoarding: keeping to remember

Hoarding was characterized by the tendency to have large amounts of digital "stuff", rarely deleting any of it. Participants often kept data even if they described it as having no value. The practice had both an emotional component (where it was a response to the fear of forgetting and letting things from the past go) and a practical component (where it was related to job or external requirements). When discussing hoarding, participants often reported challenges and frustrations with managing and "being on top" of data.

Self-identifying with hoarding

Similar to what happened with Sarah, we were surprised when some of the participants self-identified as "hoarders." "*I am a hoarder, I hoard things,*" said P17, explaining why he had a large number of ebooks. Or:

> "I consider myself a hoarder because I didn't delete them, cause I didn't clean them or delete, and I've kept all of them, except a few." (P20)

> "I'm a bit of a hoarder, I just keep all the stuff and nothing ever goes away." (P12)

However, not all participants were comfortable identifying with hoarding. At one point during the interview, P8 said: "*I am not a hoarder, really I am not!*" And added that she could delete stuff if needed. However she later explained that she "*keep*[s] *everything*" because she "*like*[s] *to keep things.*" In fact, she had digital data going back to her first computer from when she was 10 years old—she was now 25.

Lots of data, often spanning years

The first point that characterized hoarding tendencies was the large amount of data participants had kept through the years. P12, for example, had a large number of old files on her computer from decades ago that she never looked at and was surprised to occasionally discover. She also had a large number of pictures on her phone and a lot of unrecognized documents on Google Drive. P17's ebooks were in the thousands. P3, who had recently taken a trip around multiple countries, had around 6,000 pictures just from that one trip on a hard disk, which admittedly was "*a lot to deal with.*" Although they had kept it for long time, participants often dismissed most of their data, describing it as not needed. For example, P23 had four external hard drives in which she stored videos taken at public events such as concerts or festivals. She had kept them all since the 1990s:

> "I've always kept them, I know I don't need them anymore, but [I just keep them]. I guess I hoard things. At least with data it just takes the

> drives. It's not like it accumulates or takes the space in your room. Before I used to [go] shopping to buy clothes and clothes would add up. And then, OK, I'm running out of space! Get rid of the old stuff, right? So I kinda stopped that now. But I guess I switched over to data!" (P23)

This hoarding tendency did not seem limited to videos: her phone had multiple screens full from top to bottom of application folders, each with several applications. However, she reported regularly using only a few.

Sometimes participants even went as far as describing the data they had kept in rather uncomplimentary words; for example P12 said: "*Crap. All kind of stuff. My bills, recipes.*" She did not know why she had kept it all through the years: "*I don't know, might need it.*" At the end of the interview, she asked if other people did the same: "*Are there people who don't have tons of crap on their devices? Do you get rid of your messages? Do kids do that? Do kids get rid of everything?*"

Rarely deleting data

The large amount of stuff participants kept might be explained by the tendency to avoid or to rarely delete data. Participants lamented the effort it takes to curate and delete data: "*I don't think anything is going to go. I'm just going to add more, because it costs so little to add stuff but it takes a lot of time to sort the stuff you want to delete.*" (P2)

At the same time, having access to larger storage space than in the past (whether on hard drives or in the cloud) tilted the choice towards inaction:

> "I can just put it there and forget about it and don't have to actually select. If I couldn't backup to a physical hard drive and I could only backup to the cloud with a limited capacity, that would force me to clean up a little bit of the files. But because I have plentiful storage space, I don't think about it too much." (P3)

In cases when storage became an issue, getting additional storage appeared to be the easiest solution. P9, for example, described regularly buying new hard disks to accommodate her growing set of data: "I *think the reason I have my second* *hard disk is because the first is filling up, because I don't like to delete stuff.*" P8 explained a plan to keep everything in the future: "*Oh, I'll keep all of it! Well, I'll have to get a bigger hard drive* [...] *And if it doesn't fit, I will use Google Drive again if I run out of space on Dropbox.*"

The emotional value of hoarding

The costs associated with curating data in the first place might explain why participants rarely deleted data. However, this tendency also appeared to be an emotional response to the underlying fear of letting things from the past go and forgetting, a sentiment that participants often brought up. "*I like to keep memories. I don't like letting go of things*," (P8) "*I tend to keep everything. It's more like, I don't want to forget things that have happened to me in the past,*" (P11) "*I have not learned how to let go of things*." (P17)

In fact, while in some cases participants described their data as having no apparent, concrete value, it had a deeper, emotional value. Such is the case with P15, a stay-at-home mother of two, who had over 20,000 pictures on her laptop:

> "I'm sentimental. As a mom, both my children, 15 and 18, they encapsulate memories. And sometimes it feels I have to hold on to those because that's all I got left in some sense. Sometimes it feels like that. So the pictures represent something that's important to me, that's precious. The experiences with my children. [...] There's maybe this impression that things that are digitalized are somehow permanent and maybe it's an attempt to try and hold onto things, in spite of the passing of time."

Here hoarding was a proxy to remember life. It provided emotional support, with the large amount of stuff representing a large amount of experiences to go back to.

The practical value of hoarding

Along with an emotional value, hoarding tendencies also had a practical component, related to job requirements or external factors. For example, P14 reported keeping all tax documents for the previous five years, to comply with government regulations. P3 explained that a large number of pictures from a trip could act as a record for other people when looking for a job:

> "We really value these pictures, they're useful for us to keep as memories and also for employment. When they say 'Why did you have this 9-month gap in your history?' We can say 'This is what we did,' this [the travel pictures] is proof I wasn't somewhere else." (P3)

Another participant, a student in architecture who was close to graduating, explained keeping Autocad files of all of her school projects because they might come in handy when looking for a job:

> "Much of the stuff is school work. When we want to apply [for a job], make a portfolio, I've heard they ask you to send Autocad for specific projects [...] I will need a job after I graduate." (P9)

Keeping all files offered assurance that she would have the right piece of work to show when the right moment came. However, P9 was frustrated by how increasingly challenging this practice was: "*I think it's not very efficient: files are getting bigger and bigger, but my hard disks aren't, except if I get more hard disks.*" (P9) She was not alone in expressing frustrations with hoarding.

Challenges when hoarding

The large amount of data that characterized hoarding tendencies often led to frustrations. Participants reported issues in 1) keeping up with their data because of how much they had, 2) knowing what exactly they had, and 3) knowing where they had stored it. For example, P15, who valued the 20,000 pictures stored on her laptop for their emotional role, described also being overwhelmed by the sheer amount: "*It's hard to keep on top of. I wish I was more organised in the beginning 'cause now it's overwhelming going back and organising things* [...] *On one hand, you can take 30 pictures and have one that's good, but 30 pictures take time to go through.*" She aspired to become more organized and minimal in her approach, to reportedly *do things right* [288]: "*I'm hoping that by organising I can get rid of* *things, then I have more space. I hope it will be more efficient, so that whatever I have, I am valuing it and enjoying it.*" The problem was, she did not know how to go about changing her approach.

Hoarding in relation to data organization

The large amount of data also made it difficult to know what exactly participants possessed and where it was stored. This did not appear to be an effect of general disorganization. P6, who was in general methodical with her organization, had six different Google accounts that she used in the past to *segment* [293] her email usage. She also used them with Google Drive, but she had a large amount of data, so it was hard to know what it was: "*I don't know what's on everything, just random stuff.*" Similarly, P12 had a rather organized computer, making use of folders and sub-folders. Yet she had no idea what she had kept through the years, simply because it was a very large amount of data: "*I look at things and I'm like 'What's even in there?' And there might be folders inside those folders.*"

Some participants did not appear to be bothered by their approach and characterized themselves as being "just lazy", displaying a rather care-free attitude: "*Occasionally I have some weird stuff here, like this ebook, I don't know what it's doing here, this is probably my stuff from 2015. I'm too lazy to move it so I just leave it.*" (P6)

4.4.4 Minimalism: I am more than my data

Minimalism was characterized by the tendency to keep a small amount of digital data. Participants used both preventive and reactive strategies to keep as little as possible: they set a limit on the amount of data to acquire, or they regularly went back to cull it and delete it. Participants described minimalism as a way to be in control of data and life, but they also hinted at underlying anxieties behind it, and in some cases they felt detached from their data.

Prevention: limiting the amount of data

Similar to what happened with hoarding, some participants were explicit in calling out their minimalist approach. As an example, P19 had recently switched from a Macbook to a Chromebook, which she found cheaper and more "basic". She explained how the change affected her data practices: "*I am more of a minimalist now. Really keeping what I need.*" (P19) Her minimalist approach encompassed several types of data, including, for example, mobile applications on her iPhone: "*My phone, again, minimalism. I do not like having tons of apps. And the apps I don't really use, I put them here* [a folder]. *But other apps, I was so happy when they* [Apple] *said you could get rid of them.*" (P19) Interestingly, the exception to her minimalist approach was a collection of articles from the "New Yorker" magazine: "*I am obsessed with The New Yorker, the magazine. I have all different sections of it. Every time, I download it and then I read it. And I save the ones that are amazing and I want to re-read in the coming years.*" (P19)

With minimalism, some participants limited the amount of stuff to keep in the first place, and this worked as a self-imposed preventive measure: "*I try not to have too much stuff here* [the desktop] [...] *And I try not to download too much, 'cause it is primarily for school.*" (P13) Referring to her pictures, P6 explained that she was selective and therefore chose to not use automatic uploads in the cloud: "*Most people auto upload them to Google Photos, but I don't, because I don't want to save every single photo.*" (P6) It is interesting to note that P6, outside of her pictures, displayed a tendency to keep a lot of data.

Reaction: cleaning up data

Another recurring behavior participants displayed was going back to the data so that they could cull it, clean it up, and delete it as needed: "*Every couple months I go through all the old photos and delete them.*" (P21)

They articulated a thoughtful process of evaluation based on future utility and personal values:

> "With my phone, I guess I tend to only keep things that I think will be useful. For example, if I went out and took a lot of photographs in a single day, in that evening I might clean up the photographs that I didn't like or that I wouldn't think would ever be of interest to anyone else. If I don't like them, I don't think others will, and I don't see the point of keeping them. I'm generally quite clean with what I do." (P22)

In some cases, getting rid of things was the ultimate goal of being organized, an activity participants sought out: "[I] *organize, so that I know what to get rid of.*" (P19) But while some participants were very organized, this was not always the case with minimalism. For example, P21, who had a minimalist approach with the data on his phone, said he was "not an organized person whatsoever," relying entirely on the automatic organization his iPhone provided. Similarly, P16 did not have many documents on her laptop and she stored them only on the desktop: "*I have a tendency to keep my stuff on the desktop, all the time. It's not a good habit in terms of organization but* [...] *It's there, it's easy to find.*" Having a limited amount of data might have made it possible to be less organized and still be able to use it efficiently.

Underlying anxieties in minimalism

A minimalist approach was often described as a habit: "*I clean it out regularly. I don't really know why, just a habit I guess. It seems kind of busy. So, I like having clean files I guess.*" (P10) However, participants with minimalist tendencies sometimes displayed underlying anxieties behind their approach that we did not see reflected in hoarding.

Curbing the amount of data with preventive and reactive actions appeared to be a way to have control of one's data and, by extension, life: "*It's probably a way for me to stay clear.*" (P13) In some cases the need to be in control extended beyond data:

> "For me, being able to see on Gmail that I have less than a hundred emails that are unread and not having to rely on too many apps, it makes me feel calm inside. I do not like clutter. Clutter? I hate clutter! Visibly, physically, I like clean, I like washing clothes, I like seeing everything clean on the table and house. I'm not like a clean freak, that's my mother. I'm somewhere in between." (P19)

When talking about minimalism participants also expressed a need to limit the time spent with technology, displaying a general avoidance for it: "*I really don't like how much time I have to spend on it. I would rather be not staring* *at the screen for hours.*" (P13) They placed greater importance on face-to-face interactions, as if technology was in itself negative: "*I like spending time with people one on one, talking, I don't like chatting.*" (P19) This is an attitude that did not surface in participants with stronger hoarding tendencies.

Some participants also reported worries about external factors, such as money. For example, P16 used a very old computer and a four-year old iPhone, because she was trying to be economical and have a rather frugal lifestyle. She also was in a phase of her life where she did not have large amounts of data in the first place: "*I try not to do a lot. I don't have to do a lot of documentation for school, I'm done with school, so I don't have a lot of essays.*" (P16)

Similarly, P21 had recently "downsized" his digital life: he went from owning a Mac computer to having just an iPhone for all his data. This change, that "did not come from within" (implying again that minimalism was in part a reaction to financial constraints) imposed a limit on the amount of space at his disposal:

> "[When] I had more space, I would save almost every stupid photo to the computer and have photos of my background, have photos of my thumb. And now with limited space I have to be more choosy [...] I never needed all those photos [...] I do enjoy this [the iPhone] because it simplifies everything a little more." (P21)

The exception to his approach were texts. P21 explained the need to keep all of them because they were important for work and having a record of what people said.

Detachment from data

Minimalism sometimes translated to a level of detachment from the data itself, to the point of being at ease with the possibility of losing it. P21, for example, related how his approach evolved after downsizing:

> "After awhile, you know, they're just photos. And life is ongoing really. There was that big need before to hold on to every little type of thing. And now, you know, it wouldn't be the end of the world if I lost these things." (P21)

External factors such as money appeared once again to be important in determining the contextual value of data. This was the case more with minimalism than hoarding:

> "I would rather not [lose it] obviously, but I don't think it would be that critical. I would get over it pretty quickly [...] I would go to some lengths to get it back, but if it was to cost me some money, I'd rather lose the files than money. Money is more important to me I guess." (P22)

That is not say that in minimalism data did not have value. Participants reported how the limited data they kept was a part of themselves: "*The things I use more frequently are in this file. This is my D&D, I play Dungeons and Dragons. It's a big part of who I am.*" (P22) However, they also reported being at ease with the possibility of losing data. As P16 summarized: "*That's OK, if I lose it, I lose it.*"

4.5 Discussion and implications

In introducing a spectrum of tendencies, questions about their nature arise. Are you innately more aligned towards minimalism or a hoarding? Can you move across the spectrum? Can you embody aspects of both extremes at the same time?

4.5.1 Tendencies across a spectrum

We start by addressing terminology. We talk about a spectrum of tendencies with hoarding and minimalism at two extremes, rather than categorizing participants as either "hoarders" and "minimalists." This is because we saw variation both *across* and *within* individuals, and also across data types. P21, for example, had a minimalist approach with most of his data because of external factors: once he sold his computer, he became choosy with what to store, except with texts. Similarly, P6, who was highlighted in both the hoarding and minimalism sections above, displayed tendencies on both extremes throughout her various devices, hoarding the majority of stuff, while also displaying exceptions for specific types of data (e.g., photos). P19, with the strongest minimalist tendencies, displayed an exception in collecting articles from the New Yorker. Several participants shared similar patterns of behaviors, suggesting that the tendencies were context-dependent and not a clear-cut binary. Therefore, our goal was to categorize behaviors across a spectrum, not individuals.

4.5.2 Individual variation is common

A growing body of literature shows how people might segment their digital data into multiple mental places: an account for work stuff and one for personal stuff [293]; a messaging application for friends, one for family [207]. This mental segmentation adds to the idea that there is variation within an individual: users approach data differently depending on the context they build around it. Therefore, a single user can actually incorporate multiple behaviors, influenced and dependent on the specific context she needs to manage at a specific time. Chapter 6 further explores this idea.

4.5.3 Comparing and contrasting hoarding and minimalism

As two ends of a spectrum, hoarding and minimalism appeared to be radically different opposite approaches. But while there were indeed differences in required effort, they both served a similar function in helping participants construct their identity.

Identity construction

Tendencies across both hoarding and minimalism appeared to often have the implicit goal of providing participants with a framework for building their identity. This is a practice closely tied to data preservation [66, 68, 149, 211]. Participants looked at themselves in relation to data, context, and other people. Here we refer to the several quotes where participants asked what other people did with data compared to them: "*Do you get rid of your messages?*" (P12), "*Do most people have a lot of stuff?*" (P13) Similar questions were a common occurrence during the interviews. Towards the hoarding side of the spectrum, the large amount of data appeared to provide emotional support against underlying worries and concerns of time passing. Data was a symbol of experiences and memories (*I have data there-*

fore I am). On the contrary, limiting the amount of data in minimalism seemed to provide a way to gain independence from technology and detaching from data (*I am more than my data*, paraphrasing P13).

Costs and effort

Tendencies at both extremes came with costs, although at different stages of the preservation process. Hoarding tendencies seemed to have no upfront costs (e.g. P3: "*I can just put it there and forget about it*"), but later revealed themselves to not be an optimal preservation strategy if the amount of data became too large. Hoarding was a way to offset any upfront costs. Minimalism, on the contrary, required both an initial investment and ongoing dedication: setting a preventive limit and regularly going back to clean up data. In short, minimalism required ongoing effort, while hoarding seemed to require effort only once problems started arising, if at all.

Past clues about hoarding and minimalism

Previous studies contain clues about the notion of a spectrum of behaviors falling between hoarding and minimalism, but they rarely use these terms. For example, Spurgin [263], in studying photographers, talks about how some people delete all pictures, some do not, most fall in the middle. Henderson [126, 127] talks about filing and piling (two common strategies for organizing documents [177]) and mentions participants who self-identify as hoarders. We also see similarities between minimalism vs. hoarding and cleaning vs. keeping email [117], where participants either cleaned their inbox or let messages accumulate.

All these different categorizations of individual differences are neither in conflict nor duplicates. What we provide is a broader and more comprehensive lens on user behaviours that builds on top of and extends previous categorizations.

By focusing on a broad range of data types, we provide a broader context for previously reported behaviors that come from studies in specific, narrower settings (e.g. personal documents, email). That is, the spectrum of tendencies that we developed appeared to encompass several types of data, suggesting that it represents an overarching phenomenon not limited to a specific domain. We focused on data preservation, but we speculate that these tendencies might play a role in other user behaviors (e.g., tab usage in browsers or notification management). Further, we believe that looking at the different prior categorizations together in relation to hoarding and minimalism might lead to building an even more comprehensive and exhaustive spectrum of data related behaviors.

4.5.4 Reflecting on hoarding and forgetting

The tension in using the term "hoarding" is that the word itself often embodies negative cultural connotations, evoking images of people buried alive by their possessions. This might explain, for example, why P8 was so emphatic in saying that she is "not a hoarder!" But we saw how hoarding tendencies had an important role for participants, providing them with an emotional support for the fear of forgetting things, an insight that further supports the link between digital possessions and their role for identity shaping. Kaye *et al.* [149], for example, titled their paper about personal archives "To have and to hold," highlighting the importance of holding onto to things.

It is interesting to compare the emotional need of never forgetting to recent neuroscience studies about memory. Researchers suggest that forgetting is in fact a useful function of the human brain, essential to make decisions [135, 242]. Other studies show how taking pictures of every moment does not actually help in remembering them [129, 268]. There are even specific circumstances (e.g., the breakup of a relationship) where disposing of digital possessions is seen as a necessary act to avoid negative emotions [247]. Considering these insights, a worthy question to ask is whether attempting to store and keep *everything forever* still allows the space for forgetfulness and how?

4.5.5 Implications for shaping technologies

Leading technology companies such as Apple, Dropbox, Google, and Microsoft have an interest in encouraging users to move their data onto cloud platforms and accumulate large amounts of it: the more data, the more space users need. The more data, the more possibilities to thrive as a platform [264]. Unlimited storage for pictures on Google Photos might seem a generous offer, but generosity is not

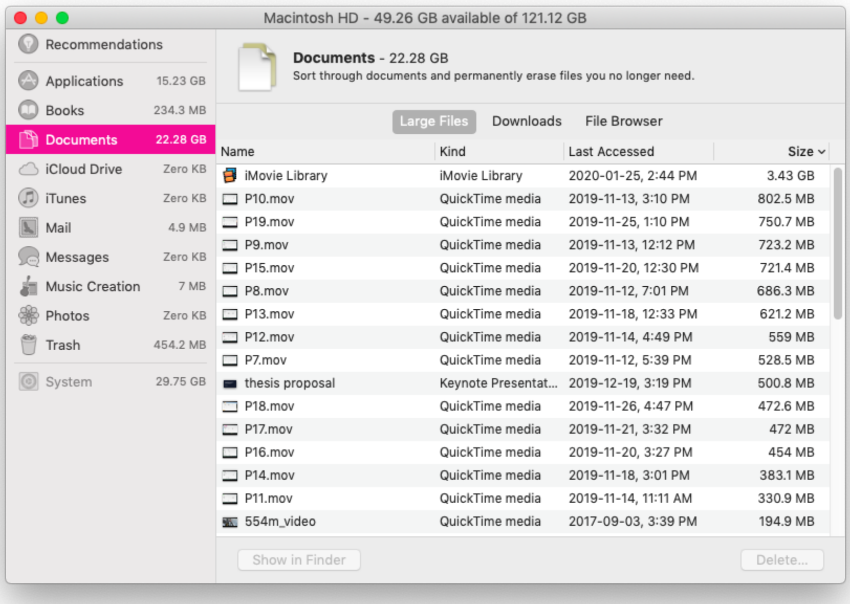


Figure 4.2: macOS Sierra shows users large files on their hard drives, displaying the size and the last time they accessed them.

necessarily the main motive when considering a larger business model where data is an essential resource for machine learning and AI training [170].

In this context, how much do technology applications influence data preservation behaviours? Some participants mentioned the amount of storage at their disposal as a decisive factor for keeping large amounts of data, while others were accumulating independent of it. At the same time, some participants gravitated towards a minimalist approach because of the limited storage on their devices. So we do not have a definitive answer to our question, but we do believe that considering the spectrum of tendencies we present can inform design decisions. What we offer are not specific design recommendations for user interfaces, but rather broad implications that could help shape technology.

Seeing these tendencies as living on a spectrum with two ends lead us to advocate for ways to mitigate the costs that characterize both sides. Some recent

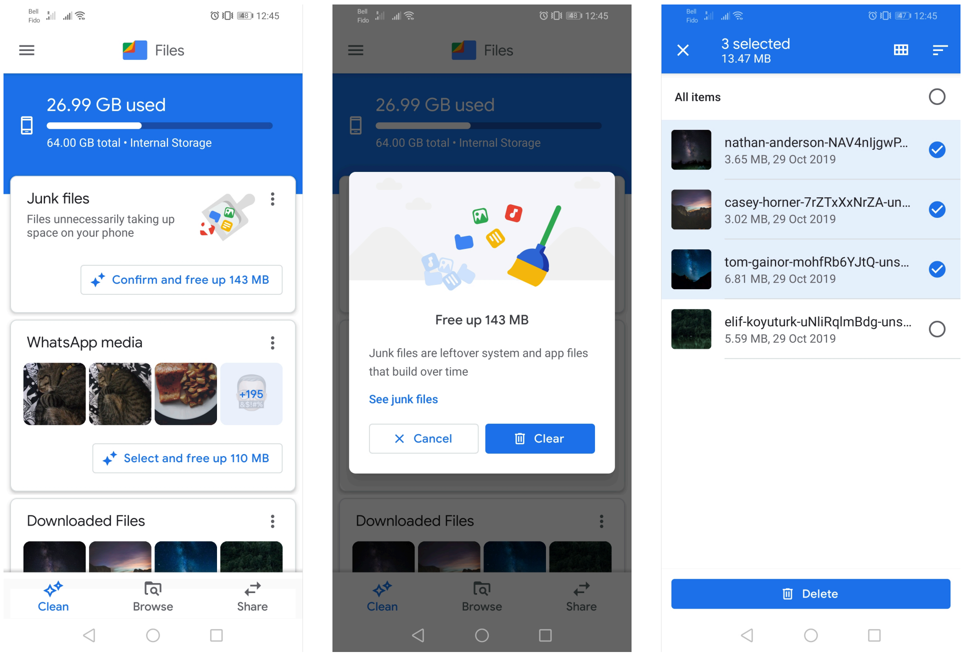


Figure 4.3: Files is an Android application by Google that gives users recommendations on how to free up storage space on their phone.

changes in interfaces show that mitigation is possible. For example, recent versions of Apple’s macOS provide a panel, although rather hidden, to explore how storage space is used (Figure 4.2). It shows what are the largest files on a user’s hard drive, their size and the last time they were used. This is information that the operating system can easily access and can be helpful to users to inform decisions about data preservation. Similarly, Google released at the end of 2017 “Files” [4], an Android application that suggests to users how to free up storage space on their mobile devices by deleting, for example, old apps and temporary files (Figure 4.3). Even though it is not clear how many users are aware or regularly take advantage of such features, their existence provides some evidence that companies are at least somewhat conscious of the frustrations experienced with the accumulation of large files.

[4]https://filesgo.google.com

We recommend more user support along these directions, namely, finding ways to increase awareness for these features, and making them more visible during daily usage. In Chapter 7 we will explore similar design concepts to understand how people might perceive them.

4.6 Summary and conclusion

We have shown how participants approached digital data preservation driven by a spectrum of underlying tendencies with two extremes: hoarding (where they accumulate large amounts of data, sometimes considered useless, experiencing in some cases challenges with managing it) and minimalism (where they try to keep as little as possible, preventing or reacting to data as a way to be in control of it). There was nuance and variation within individuals, but tendencies close to both extremes of the spectrum appeared to be a way for participants to build their own identity in relation to data (*I have data therefore I am* vs. *I am more than my data*).

The contribution and value of this study lies in: 1) bringing to light a spectrum of tendencies with hoarding and minimalism on two ends, characterizing them in depth, 2) comparing and contrasting different user behaviours, showing their common role for identity construction, 3) putting them in context compared to previously reported behaviors in the literature.

Our analysis sheds light on possible behaviors for preserving digital data, a generally under-unexplored topic. Furthermore, they have broad implications for shaping technology, opening rich possibilities for future work. Now that people are in the foothills of a new world where seductive cloud storage is pervasive, it is critical to understand what drives people's behaviors so that we can shape this world in a way that promotes informed decisions and well-being.

In the next chapter, we will complement Study 1, with a mixed-methods study that looks more closely at *decluttering* data rather than preserving it. Then, in Chapter 6, we will tease out the variation in behaviors across the spectrum and introduce a set of "behavioral styles" that expand more general preservation tendencies.

Chapter 5

Study 2

Unpacking Decluttering Criteria and Practices

In this chapter, we present Study 2, a mixed-methods study (7 interviews and an online survey with 349 participants) that we conducted as a followup to Study 1 (Chapter 4). Here, we focus on *decluttering*, defining criteria and practices that participants used to declutter their data. These criteria and practices inform the design concepts that we explore in Chapter 7. In addition, we later use the 7 interviews from this chapter as part of the overall analysis we present in Chapter 6. We also build a taxonomy of data types that will play a role in the prototype system we introduce in Chapter 8. Similar to how we used the term preservation in Chapter 4, we use the term decluttering to indicate specific actions that are part of the *keeping* stage of personal data curation.

5.1 Introduction

Data curation involves both keeping and discarding data. After looking at preservation tendencies in the previous chapter, we now take a closer look at how people decide what to *get rid of* when curating their personal data. As we have established before, personal data curation is the overall process of keeping data and managing it for future use. Past work tends to focus on the items that people *keep*, as we outlined in Chapter 2, pointing to criteria that people might use. However, as we have seen, curation is not only about keeping. It is also about discarding. Less is known about criteria for discarding data.

In this study, we aim to fill this gap by asking: what are the types of data that people consider when curating, and what are the criteria and practices they use to "declutter"? To answer this question, we conducted a mixed-methods study consisting of contextual interviews with 7 participants as a followup to Study 1 (23 interviews on "hoarding" and "minimalism", see Chapter 4), and an online survey with 349 participants. The goal of the study was to investigate "decluttering" practices with two different samples (narrow and specific in the interviews, broader and more general in the survey) to generate insights that can inform the design of personal data curation tools.

We define *decluttering* as the act of removing or reorganizing data. We see it as part of the overall process of selecting what data to keep or discard within curation. Decluttering can involve actions from across the curation cycle such as deleting, hiding, archiving, moving, and so on.

Using data from both the interviews and the open-ended responses from the survey, we outline a taxonomy of data types and decluttering criteria, together with a set of decluttering practices, that we use to inform the rest of the thesis work.

5.2 Methodology

To answer our research questions, we used a mixed-methods approach: first, we conducted 7 in-depth contextual interviews; then, we ran a broader survey with 349 respondents. We present details about each method, followed by the combined analysis.

5.2.1 Contextual interviews with self-identified minimalists

In the first phase of the study, we conducted interviews to investigate the process of decluttering possessions, both physical and digital. Our goal was to identify decluttering practices to use as a basis for design.

Participants and research process

In the interviews, we took a constructivist grounded theory approach [56], and used theoretical sampling to recruit seven self-identified "minimalists" (2 women, 5 men) from a local "Minimalist Meetup" group in Vancouver, Canada. We chose this specific population to expand on Study 1, to complement additional recent literature that focuses more on a "hoarding" perspective [271], and to better explore variation in attitudes among people who take a minimalist approach. Recent work shows that "minimalist" participants can provide inspiration for design [58, 59].

Six participants identified as Caucasian, one as Asian. They all used a computer and a smartphone, except P1, who used a flip phone. P6 also used a tablet. Their occupations and ages varied: P1 (31, receptionist), P2 (36, investment analyst), P3 (19, barista), P4 (35, system analyst), P5 (32, administrator), P6 (49, project manager), P7 (36, life coach). Their living arrangements also varied: most lived in studios or one-bedroom apartments. One lived in a minivan, one in a family house.

The research process was iterative and spanned four months. Data analysis informed data collection. After each interview, we transcribed the audio verbatim and did a first round of open coding. We proceeded until theoretical saturation. We also used data from the survey as a complement. In most cases, interviews took place at the homes of participants (Figure 5.1) and lasted approximately one hour each. One interview took place at the participant's place of work for scheduling purposes. Another interview took place in part in a minivan and in part in a park near the minivan (the participant's current living space). Participants were compensated $40.

Interview questions

The interviews were semi-structured. In all interviews we asked participants to give us a tour of their place and their possessions. Then, we asked them to recall a recent time in which they decluttered possessions and if possible to show us the process. We discussed what clutter meant for them, what was something cluttered in their home, and how they approached minimalism. We asked similar questions when switching to digital possessions. Participants showed us the data they were comfortable with, walking us through their organization and decluttering practices.

Figure 5.1: In Study 2 we ran contextual interviews with 7 self-identified "minimalists," asking them to show us and discuss how they curated their possessions, both physical and digital. In this figure, some of the homes and highlights that participants discussed (from left to right): a small but cherished apartment; a collection of ebooks contrasted with a collection of physical magazines; a frugal apartment with no visible devices; a set of boxes for "purging" items.

We also probed the difference between the two domains and any challenges they face in either or both. In the later interviews we also asked about digital tools to delete data (e.g., Clean My Mac) and whether participants used them.

Data analysis

We started the analysis with line-by-line open coding using mostly *in vivo* codes[1]. To keep the focus on actions, events, and the underlying process, we used "active codes" with gerunds [56]. We used constant comparison to better understand different contexts and individual attitudes. We later grouped codes into categories. Then, we did a second round of focused coding, and finally a round of selective coding. We also used memos: some became part of the categories we present, others come up as choices in the research process (some examples are available in Appendix C).

5.2.2 Online survey

In parallel to the interviews, we conducted an online survey about data management behaviors and decluttering episodes. Our goal was to complement the interviews with a larger sample and collect a broader range of decluttering experiences.

[1] "In vivo" codes are based on participants' words. They are a way to be alert about participants' language and preserve specific meanings [56].

Survey participants

We recruited survey participants through mailing lists, a university paid studies board, word of mouth, and online postings in Vancouver, Canada. In total, 349 participants took the survey and answered at least one question. 334 participants reported their age: 18-24 (45%), 25-34 (34%), 35-44 (11%), 45-54 (5%), 55-64 (2%), 65-74 (1%), 75+ (1%). 335 reported their gender identity: woman (69%), man (27%), transgender (1%), other (2%). Out of 335, 30% had experience with computer programming. 334 reported the highest level of education: primary or high school (32%), technical training (7%), Bachelor level (44%), Master level (16%), PhD level (0.9%). 319 reported their cultural background and ethnicity, with a wide variety (e.g., African American, Asian, European, North American, South American) 318 reported their current occupation: 127 were students at different levels; others had a variety of occupations (e.g., artist, cook, designer, nurse, project manager, teacher). Each survey question has a slightly different number of responses because only one question was mandatory. Our ethics board prescribed that participants had the right to withdraw at any point. All participants had a 1/10 chance of winning a $25 gift card.

Selected survey questions

In the survey (administered using Qualtrics), we asked three questions about decluttering digital data: 1) "When was the last time you decluttered some of your digital data, that is, you were in an active session focused mostly on deleting some of your data?" (348 responses) 2) "Can you briefly describe what you decluttered?" (325 responses) 3) "How often do you declutter digital data?" (348 responses). Question 2 was open-ended, the others multiple-choice.

In the survey, we also asked questions about decluttering physical possessions (see Appendix B), but we do not report them here because they overlap with previous work [140, 227]. Similarly, we asked participants about their general approach in keeping or deleting digital data (349 responses), their approach for different data types (e.g., documents, photos, media files, texts, apps, bookmarks, contacts, Facebook friends) (346 responses for the data types on average), and to agree or disagree with a list of statements connected to general data tendencies. We do not elaborate on these questions, because we realised they would not help us in other studies, but we report some descriptive results, together with the complete list of survey questions in Appendix B.

Data analysis

For close-ended survey responses, we report descriptive statistics based on participants' answers. To analyse the open-ended survey responses to question 2 on digital decluttering, we used open coding. We coded responses taking a mostly deductive approach, using words or sentence fragments as the unit of analysis. In the coding, we used the following categories to look for specific aspects of decluttering: what data, when, how, why, where (in terms of devices or platforms). We also developed additional inductive categories that informed the decluttering practices we describe.

5.3 Results

In this section, we present results from both the interviews and the survey, highlighting at relevant points whether they are based on one or both methods.

5.3.1 General decluttering habits

Survey participants reported decluttering regardless of whether they tended to keep or delete digital data. They decluttered a wide range of data types with photos and screenshots being the most mentioned type: photos and screenshots (mentioned by 146 participants[2]); files and documents (107); email (64); audio and video files (53); apps and programs (33); texts and voicemails (19); system data, disk fragments, cookies, cache, logs (such as call or browser histories) (10); contacts (8); games (6); accounts (4); bookmarks (4) Facebook friends (2), reminders (1). They decluttered from phones (58), computers (40), cloud platforms (16), external hard drives (3), and tablets (2). We expand on criteria for different types below.

[2]The numbers we provide here should be carefully interpreted as indicative. Because responses were open-ended and unstructured, not every participant mentioned both *what* they decluttered and on *which device* they decluttered.

When asked about the most recent decluttering session, a third of survey participants reported decluttering digital items within the last month (30%). In terms of frequency, most reported decluttering data multiple times a year (33%). But a relatively high percentage of respondents reported rarely decluttering digital items (21%).

5.3.2 A taxonomy of data types and decluttering criteria

As mentioned above, survey participants reported decluttering a variety of data types. We grouped different data types into six macro-areas (Figure 5.2): documents, organization, communication, media, system data, logging data.

- **Documents**: files and folders, productivity documents (text documents, spreadsheets, presentations, PDFs).
- **Organization**: tasks, notes and reminders, events, bookmarks.
- **Communication**: emails, texts and messaging conversations, voicemails, phone contacts, social network contacts.
- **Media**: photos and screenshots, videos, audio (music, playlists, recordings, voicemails, podcasts), ebooks, web articles, games.
- **System data**: icons for apps or links on the desktop, applications, passwords, accounts on websites and services, temporary data (e.g., cache, system logs, cookies).
- **Logging data**: tracking data (location/watching/searching/browsing history), life-logging data (e.g., mood tracking, food tracking , sleep tracking, or money tracking) [3].

Then, we looked at decluttering criteria that survey participants mentioned using for different types of data.

[3] Although survey respondents did not mention specific logging apps, we included life-logging data together with other types of tracking data to encompass a range of technologies that HCI often associates with the study of personal data [85, 87].

Figure 5.2: A taxonomy of personal data types, derived from the kinds of data that survey participants reported decluttering. We note that even though this taxonomy covers a relatively comprehensive set of data types, it is not intended to be fully exhaustive given the nature of our methods.

- **Documents**: Empty or broken content; without a file name or with a weird file name; duplicate similar to other documents; old versions of the same document; large size; old; unused; backed up on an external location or synced on cloud services; created, shared with, or sent by other people; downloaded from a website.

- **Organization**: Old; completed; irrelevant; unused (bookmarks); unreachable (bookmarks).

- **Communication**: spam or irrelevant content; newsletters; containing sensitive information (e.g., attachment with passport, address, or credit card number); old; group conversations, outdated (for contacts); unused (for contacts).

- **Media**: Duplicates; large size; blurry photos; unflattering photos; emotional content (e.g., post-breakup); irrelevant content; disliked content; not-safe-for-work content; shared or sent by others; containing sensitive information; used in another document (i.e., as attachment); content already consumed; unused.
- **System data**: Unused; large size; old; tied to a service that does not exist anymore.
- **Logging data**: Irrelevant; containing sensitive information.

These insights on data types and decluttering criteria informed the design of the Data Dashboard prototype (Chapter 8). In particular, we used the taxonomy of decluttering criteria as a starting point for populating the automatic categories in Explore Your Data, designing the recommendations in Quick Actions, and providing default options in Settings (these are all key sections in the Data Dashboard prototype, that we explain in more detail in Chapter 8).

5.3.3 Decluttering practices

Both interview and survey participants reported a mix of practices for decluttering[4]. These practices were largely consistent between physical and digital items. Yet, as we discuss later, the digital domain offered unmet opportunities for simplification and better support.

We characterize decluttering practices based on their temporal nature (*routine*, *serendipitous*, and *triggered*) and the *selection strategy* involved (*mass decluttering* or *individual inspection*). These different practices will play a role in the rest of dissertation: in Chapter 6 we will discuss how broader temporal attitudes inform behavioral styles in data curation; while in Chapter 7 we will use selection regimes as a key design dimension to explore.

In the analysis, we also touch on the role of external tools for decluttering and key differences between physical and digital decluttering that can inform the design of data management tools.

[4]In this section, we use the label *PX* (i.e., Participant X) for interview participants' quotes and *RX* (i.e., Response X) for survey participants' quotes.

Routine decluttering

Routine decluttering happened regularly, as a habit. Participants took some time to go through their data and delete some. For example, P4 decluttered his large Google Photos library every week: "*I do it as a routine. Every week I clean up my entire Google Photos to make sure that there are no photos that I do not want. I categorize it. I create the albums and I put them in my trips, so that it's not everywhere.* [...] *If I do not this, it's scattered across the entire timeline and it's a pain to search.*" Routine decluttering could be frequent or infrequent, depending on different data types: "*I delete text conversations on a daily basis, my email inbox is always managed so that it's empty. My photos are sorted weekly. At the end of each month, I declutter files. At the end of each month I declutter friends on social media.*" (R34). Routine decluttering helped participants to mark the passage of time or a transitional period in their life: "*I finished the school term and decided to organize my school notes and delete what I didn't need*" (R160).

An important distinction compared to other decluttering practices is that intrinsic motivations drove routine decluttering: respondents, on their own accord, made some time in their everyday life with the explicit goal of decluttering data, without external pressures. A key characteristic of routine decluttering was its longitudinal nature: it took place gradually over time. This points to the importance of visualizing temporal dimensions of data, something that commercially available tools to manage data do not do well–we explore this idea with *Patina*, one of the design concepts we introduce in Chapter 7. The notion of routine decluttering is consistent with insights from previous work [151, 152, 311].

Serendipitous decluttering

Serendipitous decluttering occurred in the context of another activity. For example, when participants were busy and not planning to declutter. But, by encountering some items that fit decluttering criteria in a serendipitous way, they decided to declutter. Serendipitous decluttering could happen while browsing through data: "*I was looking through old photos and realized I could delete a bunch of them*" (R144). Or, while looking for something specific and failing to find it:

"*When I try to find something and I can't find it, I usually go through and filter out what I don't want/need anymore, which will declutter my digital data*" (R170).

The main difference between serendipitous and routine decluttering was the unexpected and unplanned, yet ultimately productive, outcome of the process. The context of items was essential to serendipitous decluttering: items popped up as candidates for decluttering in relation to a broader set. Once again, decluttering criteria were important to decide what to get rid of, but many of the criteria are essentially invisible in digital items and require careful consideration from users.

Triggered decluttering

Triggered decluttering took place after a relatively infrequent event. For example, a breakup [247], buying a new device, or space running low. For digital items, the trigger was often a notification from the operating system: "*My mac kept popping up with a notification telling me I only have X GB of storage, your disk is almost full.*" (R271) Participants felt forced to act and discard items, even though this might not have been what they wanted to do. For example, P2 had a strongly minimalist approach with physical possessions, but not so much with digital data, rarely finding the time and the will to delete things. On his 32GB smartphone, he had only 1GB of storage left. He explained that he recently had to delete some pictures because of an update: "*I deleted some photos* [because of] *the regular update. It wanted to update the OS and I didn't have the choice. The pop up message said the update will install when I restart the phone. So I had no choice, right? I had to do it* [delete]. *So I got prompted again, that's when I deleted.*"

Triggered decluttering shares some aspects with serendipitous decluttering: they both occur unexpectedly. But external factors motivate triggered decluttering while serendipitous decluttering relies on internal motivations.

Triggered decluttering was also time sensitive and often needed to be targeted towards an end goal (e.g., freeing up space). In the digital context, this can be hard, because digital items are often experienced as "placeless, spaceless, and formless" [215], and it might be hard to find the best candidates to achieve the end goal: "*My phone would not allow me to download a meditation app I wanted. It was frustrating getting rid of enough space and I didn't have a clear idea of what took* *how much space.*" (R48) But digital tools could proactively find items to delete–an idea that we explore with *Data Recommender*, one of the design concepts we introduce in Chapter 7. Previous mentions of specific triggers to declutter in HCI research include storage space when dealing with shared devices in the home [109], changes in important relationships (e.g., separation, divorce, or death) [249], and changes in social status [151].

Selection strategies

Based on how participants *selected* items to declutter, we describe two main strategies: mass decluttering and individual inspection. These strategies are consistent with work by Jones & Ackerman [140], who highlight similar selection *regimes*.

Mass decluttering, or "purging," involved getting rid of many items at once (a finding consistent with Kim [151]). For example, P6 had a "purging week" at the start of spring every year: "*This one-week period in May.* [You found me] *right in the middle of a purge.* [...] *I'll stay here, spend the whole week doing minor repairs on the house, patchwork. And the purge.*" Books were among the items he would purge; throughout the year he put aside the ones he did not want to keep so he could donate them to a public library.

Mass decluttering considered items and their merits as part of a larger group or category rather than individually. This strategy was often a way to jump-start the habit of decluttering: participants wanted to get rid of the bulk of clutter so that they could follow on with a more gradual and regular process after it. For example, this is how P4 recalled a digital downsizing he had gone through: "*I did a massive cleanup maybe two years ago before I gave my desktop away. [...] I cleaned up my photos in Google Photos and then I thought, let's clean up other data as well. I started cleaning up my contacts. In my phone, I had around 200 contacts, some people I did not even contact in years! [...] You know, things like my old dentist, my old doctor, those are just irrelevant. [...] It took long because there was a lot of stuff I had to go through and delete or save*."

A complementary selection regime involved going through items one by one through ***individual inspection***. This was a cross-cutting strategy that often required more time. For example, P1 recalled what happened when she had to move to a smaller apartment (triggered decluttering). She went through clothes and items one by one and decided what to get rid of based on their use: "*We had a very short time.* [...] *The process was really 'Do I wear this? Do we use this?' And anything that wasn't getting much wear and usage was gone*" (P1). Several prior studies also mention the need for users to inspect items on an item-by-item basis before deciding what to discard [111, 150, 151, 299].

The role of external tools in decluttering

When it came to external help for digital decluttering, only one survey respondent mentioned using in-person company support to decide what to declutter: "*I was running out of space on my computer. I couldn't figure out what was what, so I had to go to the Apple store genius bar and have them help me discern what was deletable.*" (R22) Some interview participants had heard of or had used tools that can help declutter (e.g., CCleaner, AppCleaner, Clean My Mac). But in their view, these tools only supported basic behaviors, failing to address the most relevant data: "[Those tools are] *mostly to clean up space and temporary files, probably not main files.*" (P5) In Chapter 7 and Chapter 8 we will explore ideas and solutions that can provide a more comprehensive support for decluttering and data curation.

5.4 Discussion

5.4.1 Physical vs. digital decluttering

When we recruited self-identified "minimalists" for the contextual interviews, we expected they would have a similar approach for physical and digital possessions. But this was not always the case. Several interview participants mentioned how it was more difficult decluttering digital items, a result consistent with past related work we discuss in Chapter 2.

Several qualities of digital data (e.g., spacelessness [215]), contributed to making digital decluttering a difficult task for many. In addition, the tools that participants reported using did not offer adequate support for the varied and contextual practices that they used. These results further motivate the need to explore new ways for supporting data decluttering and selection, something we do in the rest of the dissertation. Here, we outline the general design directions we derived from this study and later explored in Chapter 7.

5.4.2 Opportunities for supporting digital decluttering

Visualizing temporal aspects of data

In comparing physical and digital items, it was clear that for digital items it is not easy to get a quick, visual sense of the different decluttering criteria that people might want to refer to. Physical objects can give a better sense of them: they often show traces of use and age, for example. If a user wants to know how frequently they have used a digital item in the past month, they might struggle. Other criteria are difficult for current digital systems to capture in the first place. It is difficult for a file browser to know what data a user dislikes, or what data they have little emotional attachment to? Machine learning might offer a possible remedy to this lack of knowledge, but a more immediate solution might be to leverage metadata for visualizing temporal aspects of data. For example, data management systems could show the age of items or the frequency of interactions at a glance, with embedded visualizations. Design work in this area could leverage the idea of a digital "patina", which is recurrent in HCI literature [99, 167, 211] and has clear ties to "edit and read wear" [132]. This is one of the key design ideas we explore in Chapter 7, connecting to attitudes that different behavioral styles (Chapter 6) might have.

Proactively finding items to declutter

Several technology companies are starting to recognize some of the issues around space optimization and unwanted data that we have surfaced in this chapter. A common approach to address this problem involves creating recommender systems that can proactively help users. For example, in 2018, Windows 10 introduced *Storage Sense* [35], a new setting that can automatically delete temporary files or files in the Downloads folder that are older than a set time. Similarly, Google Photos provides users with recommendations on how to declutter their cloud-stored pictures. These examples show how recommender systems are gaining popularity in common data management tools. However, it is not clear what tensions they introduce in the space of personal data curation or how different users might perceive and react to them. Will people trust the system to know what to discard? Should these systems aggregate multiple items for mass processing forgoing individual inspection? These are key questions to address, that we later explore in Chapter 7.

Preventing unwanted data accumulation

We have seen how participants encountered many types of data that they perceived as clutter across their platforms and devices. In many cases, their decluttering practices were a reaction to unwanted accumulation of items that they did not consider important or relevant. If we take the idea of systems proactively recommending items to discard to its extreme, we can imagine systems that instead prevent the accumulation of unwanted data in the first place. Previous work has explored self-destructing data in emails [97] or having a lease on data shared with others [192]. Similar approaches could also apply to a broad range of data types and be more nuanced. A key opportunity is to explore solutions for capturing this nuance and understanding how different users might experience similar extreme tools. In Chapter 7, we will better explore the potential for this type of approach to data curation.

5.4.3 Limitations

It is important to acknowledge the limitations of this study and in particular of the online survey we conducted. In the survey, we relied on self-reported answers from participants, without being able to see their digital data. The survey sample also had one third of respondents with a Computer Science background or experience in programming. However, when we separated their answers from the rest and conducted a visual inspection, we did not notice any major differences in the descriptive results.

When looking at the survey as a whole, there can be an apparent conflation of *decluttering* with *deleting*, but that is not intended. Even though the phrasing of the open-ended question on decluttering suggests that decluttering is mostly about deleting data, participants reported actions other than deleting in their answers, although our wording might have skewed responses towards focusing on deleting as opposed to other aspects of decluttering and curating data. We decided to describe decluttering as mostly related to deleting data to keep the question simple and understandable by potential respondents: we distributed the survey in several countries, with some having English as a second language, therefore associating decluttering with deleting seemed a reasonable simplification. Terms such as archiving and curating might have been more difficult to understand for participants whose first language was not English.

Overall, we see the survey as largely generative, with its insights informing the design studies that come after it. The limitations in our questions and sample prevent us from making stronger claims about the results. However, these results can be a starting point for future research.

5.5 Summary and conclusion

In this chapter, we used data from interviews with self-identified minimalists and an online survey to outline a taxonomy of data types and decluttering criteria. We also described a set of temporal practices and selection strategies that participants used to declutter their data. These results suggest opportunities for better supporting curation in data management tools and highlight the contextual nature of user behaviors.

In the next chapter, we will synthesize our work on individual differences and user practices with a set of five behavioral styles that can bridge empirical work with design work.

Chapter 6

Study 1, 2, 3, 4

Synthesizing Behavioral Styles in Personal Data Curation

In this chapter, we take a bird-eye view of all the four main interview studies that make up my dissertation. This is an unusual approach. The five behavioral styles that we present are a direct extension of Study 1 (Chapter 4) and Study 2 (Chapter 5). But they are also informed by Study 3 and Study 4, two design studies that appear later in the dissertation (Chapter 7 and Chapter 8). We choose to present the behavioral styles early in the dissertation because they are necessary to understand the recruitment process in later studies (Chapter 7 and Chapter 8).

6.1 Introduction

In previous chapters, we explained how understanding individual differences for deciding what personal data to keep or discard is one of the key questions in PIM research [144]. We also highlighted how previous research tends to focus on individual differences in *organizing* and *retrieving* data, leaving the *keeping* stage of curation underexplored. Previous studies in HCI and Information Science report how people's digital archiving strategies are "widely varied" [259]. But despite knowing this, a systematic model of individual behaviors is missing from the literature. Identifying different types of users and their overarching approach to curation has been a long-standing unaddressed need for the past decade. In discussing a life-cycle approach to personal data curation, Williams *et al.* [309] ask whether there might exist "archetypes" for digital archiving. Gwizdka & Chignell [118] mention the possibility of "PIM personalities." Bergman [23] outlines key variables for how individual PIM behaviors differ, encouraging researchers to identify a set of overarching "PIM styles." And Khan *et al.* [150] ask for a precise categorization of "archetypes" in relation to keeping and discarding decisions. This body of work shows that categorizing user behaviors is an important step for creating and then evaluating new, personalized curation tools. In this chapter, we propose a set of behavioral styles to address this need.

Our work in this space started with the Study 1, an exploratory interview study on general tendencies for keeping and discarding data (Chapter 4). Analyzing 23 interviews with a broad sample, we identified a spectrum of behaviors with two extremes: "hoarding" (keeping most of data) on one side, and "minimalism" (keeping as little as possible) on the other. In Study 1, however, we highlighted that there is considerable variation within the spectrum, leaving space for a more detailed categorization.

In this chapter, we pick up on the premise of better categorizing individual differences and turn it into a key research question: *How can we make individual differences in personal data curation actionable for research and design?* To answer this question, we propose a set of behavioral styles in personal data curation: *Casual*, *Overwhelmed*, *Collector*, *Purger*, and *Frugal*. These behavioral styles differ along a set of behavioral dimensions and individual characteristics, while also pointing to temporal aspects of personal data curation.

To develop the behavioral styles, we used an iterative analysis process that spanned all the four studies that make up the dissertation. Study 1 (Chapter 4) was our initial exploratory study that focused on "hoarding" and "minimalist" tendencies in data preservation. Study 2 (Chapter 5) was a followup study consisting of 7 interviews with self-identified "minimalists" (Study 2 also included an online survey, but we do not focus on the survey results here.) Then, in Study 3 (Chapter 7) we conducted 16 interviews as part of a design-led exploration around data curation. Finally, in Study 4 (Chapter 8) we conducted 18 interviews for evaluating a system designed to accommodate the five behavioral styles that we had developed.

After describing the behavioral styles, we chronicle how their descriptions evolved over time with the addition of new insights, and how we used them as a recruiting tool in later studies. To ground our analysis, we reflect back on our design and research practice across the studies. We explain how we developed the first version of the behavioral styles after Study 1 and 2, and later used Study 3 and 4 to validate them [1] and enrich them. Then, we discuss how the behavioral styles help us better understand data curation, and how our specific approach to formulating them can inform research practice. We conclude with a range of design opportunities that build on top of the behavioral styles.

In this chapter, we make three main contributions: 1) We present an actionable set of behavioral styles that expand our understanding of personal data curation practices. Designers, practitioners, and researchers can use the behavioral styles in future user studies and product design. 2) We provide a reflexive account of the iterative analysis process that led to the behavioral styles. Designers and researchers can use this account to inform similar user modeling efforts in other domains. 3) We generate a set of opportunities tailored to the different behavioral styles that can drive future design work on personal data curation.

6.2 Related work

The behavioral styles we present in this chapter advance our understanding of individual differences in PIM with a specific focus on curation. In this section we outline key known differences and categorizations of behaviors in PIM to contextualize our work.

6.2.1 Individual differences in PIM

A recurring insight in PIM studies is that people manage data in different ways. Bergman proposes 15 variables that can account for differences in PIM behaviors and groups them into five categories [23]: "organization, structure, work process,

[1] When we refer to validation, we are still basing our approach on Grounded Theory procedures and criteria for qualitative research outlined in Chapter 3. Validating the behavioral styles does not mean that we are going to prove them as unequivocally true. Instead, we are checking them against new data, seeing how they hold up, and identifying any theoretical gaps in our current explanation that need iteration.

memory, retrieval." For example, organization can vary from ordered to disordered, while structure can result in large or small collections. Differences along these variables lead to different categories of behaviors. For example, common categorizations for organizing documents are piling (letting documents pile as part of shallow or "disorganized" hierarchies), filing (taking the time to organize documents into hierarchies with moderate structure) and structuring (intensive organization with deep structures) [128, 177]. For emails, cleaners and keepers are two key categorizations [117]. Cleaners tend to remove task-related documents, to-dos, and event reminders from their email inbox, while keepers tend to leave them. These strategies, however, are not exclusive and people often mix them depending on context [38, 128]. There are additional categorizations such as email prioritizers and archivers [173], or frequent filers, no-filers and spring-cleaners [306], that previous work summarizes in more detail [38, 118, 128, 220].

Work from psychology and cognitive sociology can explain why individual differences in organization exist in the first place. For example, personality traits like conscientiousness or neuroticism can help predict differences between filers and pilers [187]. People also differ in how they think of categories and these differences influence their organization style. Some people perceive categories as fixed, with clear distinctions and boundaries–their file organization reflects this way of thinking, with rigid structures that do not change over time [219]. Others do not perceive clear boundaries between categories and so they do not use any specific structure in their organization (e.g., storing all their files on the Desktop). Most people, though, might be in between the two extremes: they are flexible and can change or remove boundaries over time [219].

People also differ in how they retrieve information [25, 28, 302]. Despite improvements in search functions, users often prefer navigation for retrieving files [28] and search for retrieving emails [25]. The preference for one or the other might be a personal trait (i.e., variable between individuals, but consistent within them) [25] and might change as people age [33].

There are two key takeaways from this body of literature. One, data curation is a subjective process [22, 24, 26, 158]. Any supporting tool for different curation stages will need to rely on models of individual behaviors [23, 219]. Two, people tend to distribute themselves in clusters along *spectra* of behaviors [23, 219] (and

Chapter 4). Our work adds to the literature on individual differences in PIM by focusing on the first stage of curation, where people decide what data to keep or discard. Other than Study 1 (Chapter 4), we are not aware of models for individual differences about keeping and discarding decisions. We also look at how organization behaviors, the second stage of curation, mix with keeping decisions. On a few occasions we touch on retrieval, the final stage of curation, but we do not consider it as a key focus of our work and refer to past work for more details on this stage [24, 28, 274, 302]. In the remainder of this chapter, we use the term *curation* to largely refer to the first stage focused on decisions about what data to keep or discard.

6.3 Methodology

To develop the behavioral styles, we draw on a total of 64 interviews from the four studies that make up the dissertation (Figure 6.1). Each study had specific research questions and used somewhat different methods, that addressed the evolving nature of our overall investigation: we started with exploratory interviews to understand general behaviors (Study 1), followed by more focused contextual interviews to expand our initial results (Study 2). Then, we moved towards research through design in Study 3 and 4, where we explored design artifacts as a prompting tool for building knowledge [317]. All studies shared an interview component with overlapping questions about what data participants tended to keep or discard, general management practices, feelings and attitudes about data curation. These are the questions and answers we focus on. Our overarching methodology is in line with constructivist grounded theory [56], with theoretical sampling informing the move from one dataset to the other. After a broad interview sample in Study 1, we moved to a specific and narrow sample in Study 2, and then came back to a broad, but purposefully varied interview sample in Study 3 and Study 4. Below, we summarize the four studies. All interviews took place in Vancouver, Canada and lasted between 40 and 100 minutes (on average one hour).

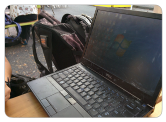
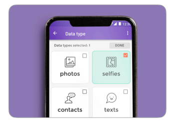
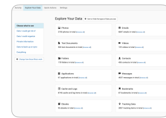
	Study 1 (2017)	Study 2 (2018)	Study 3 (2018)	Study 4 (2019)
				
	Chapter 4	Chapter 5	Chapter 7	Chapter 8
Focus of the study	Exploratory interviews (23) on how participants decided what data to keep or discard	Contextual interviews with self-identified minimalists (7) on decluttering strategies	Elicitation interviews (16) about five speculative design concepts for data selection	Evaluation interviews (18) focused on a prototype system for data curation
Interview sample	Broad sample (16 women, 7 men, aged 21-64)	Focused sample (2 women, 5 men, aged 19-49)	Varied sample (9 women, 6 men, 1 non-conforming, aged 23-71)	Varied sample (12 women, 6 men, aged 18-64)

Figure 6.1: An overview of the four interview datasets and the related studies used for the analysis of the behavioral styles. For each study, we report the year, the corresponding chapter, a summary of the general focus, and an outline of the interview sample.

6.3.1 Study 1: exploratory interviews with a broad sample

The first study (Chapter 4) took place in 2017 and consisted of exploratory interviews about general keeping and discarding decisions with 23 participants (16 women, 7 men, aged 21-64), with a broad set of occupations (e.g., consultant, cook, social worker, software tester). The interview questions touched on what data participants had kept through the years on different devices and how they decided what data to keep or discard. Participants showed us how they organized their data, discussed what they considered important to keep, and their frustrations or challenges with curating data.

6.3.2 Study 2: contextual interviews with "minimalists"

The second study (Chapter 5) took place in the first half of 2018. It consisted of contextual interviews with 7 self-identified "minimalists" (2 women, 5 men, aged 19-49), who had different occupations (receptionist, investment analyst, barista, system analyst, administrator, project manager, life coach). (The study also included an online survey, but here we focus only on data from the interviews.) The interviews took place at the home or office of participants (Figure 5.1). The interviews were semi-structured and focused on how participants decluttered data. Participants showed us their data and walked through their organization and curation practices.

6.3.3 Study 3: design exploration with with a varied sample

The third study (Chapter 7) took place in the second half of 2018 and consisted of semi-structured interviews focused on eliciting reactions to a series of design concepts for curating data from 16 participants (9 women, 6 men, 1 gender nonconforming, aged 23-71), who had a broad set of occupations (e.g., HR specialist, journalist, photographer). We recruited participants using a short description of the behavioral styles in a screening survey to recruit participants based on their different general approach to curation. The short description (included in the Results section) focused on the key aspect of each behavioral style and was largely stable before Study 3. (The complete screening survey is available in Appendix D). The interview questions touched on how participants organized and curated their data over time. We also showed them some design concepts to help them discuss data curation in more detail, reflecting on their behaviors, goals, and frustrations. The design concepts were also a way to provide support for the different needs and approaches in the behavioral style: some concepts were reflective and user-controlled in their approach to data curation and selection, others were more extreme and automated. We thought that different extremes of the design dimensions we explored could be a good match for different styles of behavior. After the study, we used the the 16 interviews to support and refine the behavioral styles.

6.3.4 Study 4: system evaluation with a varied sample

The fourth study (Chapter 8) took place in 2019 and consisted of structured sessions with 18 participants (12 women, 6 men, aged 18-64), with varying occupations (e.g., occupational therapist, sales associate, social worker). The study sessions focused on evaluating Data Dashboard, a prototype system for curating personal data, designed to accommodate the different behavioral styles. For recruiting, we used the same short description of the behavioral styles from Study 3. All the details of the prototype and evaluation are in Chapter 8. Here, we focus largely the introductory interview part of the evaluation, where we asked participants to discuss and show us their personal data curation practices. Participants talked about how they decided what personal data to keep or discard, remembered specific episodes in which they discarded or re-organized a certain number of items, and discussed any specific tools they used to curate data.

6.3.5 Data analysis

To synthesize data from the four studies we used both a deductive and inductive analysis process. Here, we describe the process linearly for clarity, but we later expand on its iterative nature and how the behavioral styles evolved over time. In all of the studies we started with inductive open coding, as part of a reflexive thematic analysis [61] (Study 1, 3, 4) or a grounded theory approach [56] (Study 1 and 2). We started by using *in vivo* codes, then grouped codes into categories, and then evolved initial categories into more abstract concepts. The inductive coding phase was useful to connect participants' quotes to specific patterns of behaviors. But throughout the studies we also used a more deductive approach, taking a high-level look at study participants and comparing them based on a set behavioral and design-oriented dimensions (Figure 6.2):

> *Quantity of data*: how much data participants dealt with on a daily basis.
>
> *Preservation tendency*: how much they tended to keep or discard, from most to a little. (This variable overlaps with Bergman's collection size [23].)
>
> *Organization approach*: how they approached their data organization (structured, unstructured, a mix). (This variable overlaps with Bergman's variables around structure [23].)
>
> *Feelings*: how they related to data and curation (attentive, satisfied, frustrated, indifferent, etc.).
>
> *Thoughts*: what were their priorities when curating data.

Pains: what were their main pains and frustrations with curating data.

Goals: what they wanted to achieve when curating data.

The first three dimensions are behavioral variables based on the variation we saw across participants in Study 1 (Chapter 4) and Study 2 (Chapter 5) but also overlap with related work on PIM variables [23]. The last four dimensions borrow from the empathy mapping framework [100]. Empathy maps are a common design tool for building user models. They focus on what users say, think, do, feel, and want to achieve during an activity or when using a product. One possible shortcoming of empathy maps is their static nature [256]. Throughout the analysis and this chapter, we overcome this potential issue in three ways. First, we see the empathy mapping framework as a starting point for coding rather than a prescriptive tool. Second, we ground our analysis in specific experiences and episodes that participants reported. Third, we elaborate on the dynamic and composite nature of the behavioral styles across time and contexts. This approach combines the practical nature of a tool such as an empathy map with the focus on narratives and user experience of more complex analysis techniques.

With these two parallel approaches, one inductive, one deductive, we focused on recurring patterns among participants using a process of constant comparison for the different dimensions. Then, we clustered the patterns in a set of behavioral archetypes. Behavioral archetypes are similar to *personas*, another common design tool for user modeling. But they differ in giving emphasis to patterns of behavior over demographics [20]. Archetypes are also more contextual than personas: a person can embody different behavioral styles over time or when changing context [20]. In general, building archetypes, or personas is a common, established step of design research [37, 238]. Outside PIM, some examples of similar efforts include studies about privacy [83], games [223, 279], or child-centric design [7].

We named the behavioral styles with labels based on feelings ("overwhelmed"), attitudes ("casual," "frugal"), or behaviors ("collector," "purger"). Because the behavioral styles share some aspects or behaviors, the labels highlight a key distinctive trait. For example, both the collector and purger behavioral styles share a disciplined attitude. But choosing a behavior as their label emphasizes the difference between them.

We also decided to label these patterns as "behavioral styles" instead of "archetypes" or similar labels for two reasons. First, the word "behavioral" emphasizes how our descriptions focus on participants' behaviors. Second, the word "styles" helps explain how to interpret them: these patterns are context-dependent, meaning that different people might have different styles in different situations or at different points in time. As a whole, the set of behavioral styles are useful to identify relevant patterns within individual participants. A real person who perfectly matches the Collector behavioral style may not exist because individual behaviors are dynamic and contextual. The styles are a cross-cutting synthesis of individual behaviors and only represent a reference point that we check individual people or participants against. Then we can say, for example, that a person or participant matches a bit of a collector, a bit of an overwhelmed, nothing of a purger, and so on. Later in this chapter, we explain the process we used to check individual participants against the different behavioral styles in Study 3 and Study 4.

6.4 Behavioral styles and approaches

As we explain above, the behavioral styles we present are a *synthesis* of dynamic and contextual individual behaviors. We see the behavioral styles as dependent on the many data types (Chapter 5) participants had to manage and curate. For example, one person might feel overwhelmed by emails, take a casual approach with work documents, and purge photos. For each behavioral style, we give an overview of key dimensions and behaviors using a high-level description and a short example. Then we discuss temporal aspects of participants' behaviors.

6.4.1 Casual

Participants who took a "casual" approach to curating their data had no particular worries, having a relaxed attitude to data management and organization. They kept some items and discarded others, with no strict rules or underlying concerns. Similarly, they did not report substantial challenges in finding data when necessary. Sometimes they described themselves as "too lazy" to do things differently. But they also thought that data curation was not an important activity in their life: in theory, they could have been more systematic about curating and managing data,

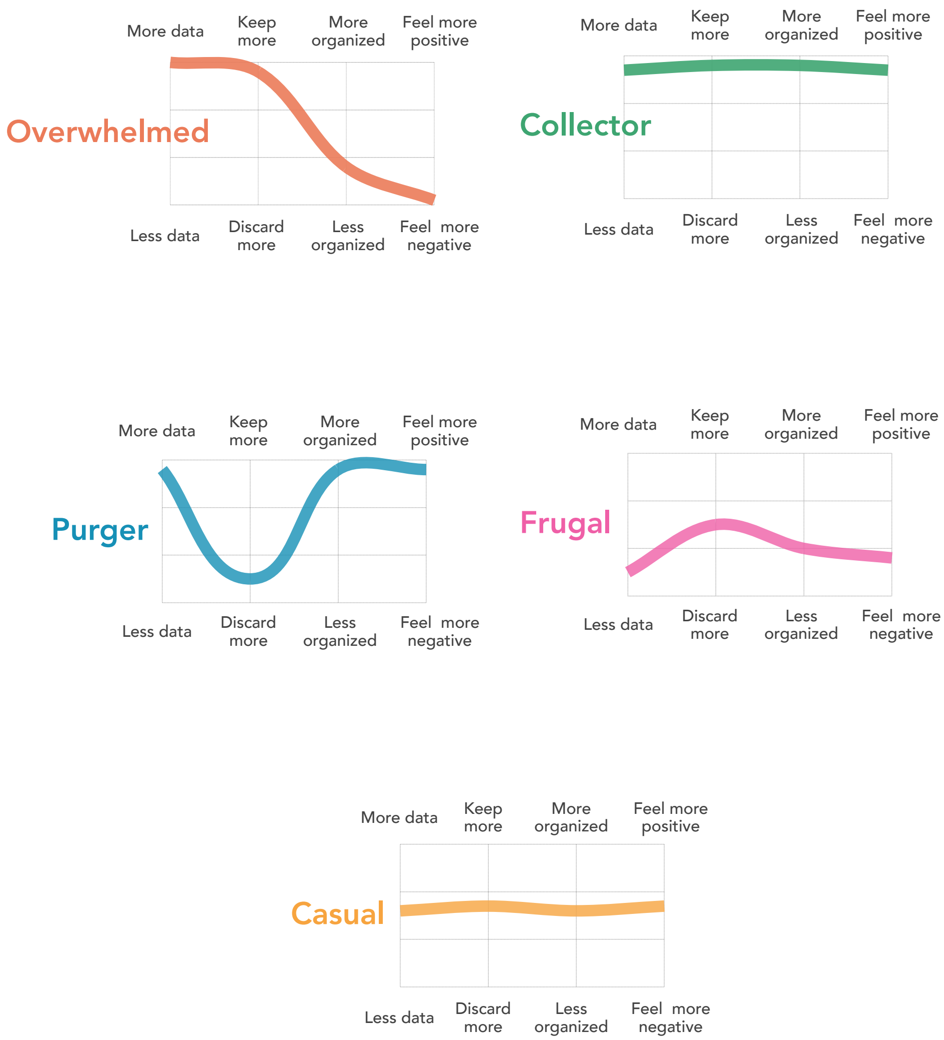


Figure 6.2: A visual overview of how the five behavioral styles map to key dimensions we used in the data analysis: quantity of data (more data - less data), preservation tendency (keep more - discard more), organization approach (more organized - less organized), and feelings (more positive - more negative). The graphs are illustrative and do not represent a precise numerical comparison between the behavioral styles.

but they did not see why they should have bothered. Their priorities were elsewhere. Even when discussing privacy and security aspects of their personal data, they did not appear particularly concerned. Some factors that influenced this approach included having newer devices with a significant amount of free space or being willing (and able) to pay for cloud storage space when needed.

For example, one participant explained that he did not take the time to curate ebooks because the time investment to do so was not worth it:

> "My digital books [...] I probably have hundreds of epubs and PDFs in here, some of which I read, some of which I haven't and yeah, I would want to go in here and delete them in principle, but in practice I don't. I don't have six to ten to twenty hours to do this just to save a few megabytes of storage. It doesn't bother me that these are here. [...] I am not going to get to this. I have more important things to do." (P6, Study 2)

6.4.2 Overwhelmed

Participants who felt "overwhelmed" tended to acquire or create large amounts of data and keep most of it. However, they did not feel positively about their approach because, in fact, it was not a fully intentional approach and they wanted to change it. Often participants explained how they tried to organize and curate data, but they failed because they did not have enough time to keep up with the amount of data, or they did not have the knowledge required to organize as efficiently as they would have liked to. Participants could be too busy with other responsibilities in life and their interactions at work or in personal life often forced them to deal with large amounts of data on a regular basis, with more data coming in than they were able to deal with. Resigned to accumulating data, they experienced challenges in knowing where to find their digital items or how to better manage them. Several participants who felt overwhelmed thought that proactive, automatic tools could help them overcome their challenges in curating data.

As an example, this is how one participant described feeling "overwhelmed" and resigned when thinking about data from many apps, touching on challenges in the whole curation process, from keeping decisions, to organization and retrieval:

> "Sometimes [it is challenging to] just remember where I saved [an item]. Sometimes when I'm looking for a file–I'm a very obsessive note taker. I have a lot of notes apps, I use Evernote, Google Keep, Simplenote. I used to have a lot more notes app but I found it hard to back it up. I remember something [and think]: 'Where did I save it? Where did I put it?' because I have it in several places. [...] I just don't feel very productive. Sometimes I feel overwhelmed by all this stuff that I keep track of. I think if there's a system, a better way to keep track. I keep a list of all the things I want to do in the future, but I feel there's not a way I can organize it better." (P10, Study 4)

Another participant complained about "being behind" with her data management and emails in particular, due to a lack of time after having a baby. She wished to return to a more manageable state:

> "I'm actually behind on my emails as well, what a surprise! So this for me [Gmail inbox with a list of emails] is not a good example, I like to have maybe one page of emails, my inbox is my to-do list. This for me is too much, I like to have five emails normally, and then I put everything in a folder, I have quite a few folders, and I've always worked like that. [I haven't taken care of them] purely [because of] time." (P7, Study 2)

6.4.3 Collector

Participants who took a "collector" approach tended to acquire or create large amounts of data and keep most of their items, often using multiple devices, cloud accounts, and external hard disks. In terms of amount of data, they were similar to "overwhelmed" participants. But in this case, participants were happy about their approach, sometimes showing pride in their collections, and did not want to change it. They reported being generally organized in how they managed data and feeling "on top of things," similar to "purger" participants. They felt that curating data was a personal responsibility and were somewhat skeptical of tools that could support this process. They took many actions to make sure that they could preserve and access their data in the long-term, for example, by doing regular backups and archiving items in multiple locations. Their main worries were about the possibility of losing data (either from a device or in a data leak), the attention required to "sync" data across platforms and devices, and the effort required to manage space efficiently so that they could "keep everything."

For example, one participant described choosing devices with large storage capacity to hold data collected over the years:

> "I get electronics that have a capacity to hold things because I tend to store stuff. I have a 3TB hard drive where I keep my photos. But I also keep them on my laptop, on my phone, all the time. I bought a 128GB [phone] because I have a lot of data that goes and comes and I don't like to delete it all the time. I do a cleanup of my phone and laptop maybe once a year and that's only because I get a message like, hi, you need to clean your phone because there is no more space." (P9, Study 3)

6.4.4 Purger

Participants who took a "purger" approach tended to only keep necessary data, even though they often had a relatively large amount of data to manage. They reported regularly "cleaning up" devices and described themselves as being quite organized. Curating and upkeeping personal data was an important activity for these participants, who spent considerable time decluttering and reorganizing items on a regular basis. Sometimes the limited amount of space on their devices drove their behaviors. But often, they reported purging data because of reasons other than space: for example, they wanted to access data more easily, they disliked keeping unnecessary data, they wanted to protect privacy, and they enjoyed the activity of cleaning up data in itself. Overall, they felt positive about their approach, often saying that they were "on top of things" and they wanted to continue managing their data in a similar way. In fact, they were interested in tools that would make them more efficient and were often open to the idea of optimizing curation through automation.

One participant, for example, described regularly cleaning up data to find items faster saying they were "a bit OCD":

> "Every two or three weeks I delete those files I don't need. [...] If I put a file in Google Drive and I try to find it, I don't wanna scroll down to everything and check the names. [...] For things I don't need, I'm like, why keep it? Yeah, I don't want to keep [unnecessary things] there. I just don't want to. It's kinda OCD. For the phone I periodically delete pictures I don't want and apps I don't use anymore. But on the phone it's partly due to storage problems. Which is not that much of a concern with my Mac. But in general I just don't wanna keep useless things." (P3, Study 3)

6.4.5 Frugal

Participants who took a "frugal" approach tended to keep little of their data, but, unlike other behavioral styles, they did not have a lot of personal data to begin with (for example, because their job or personal life did not require them to deal with data frequently). These participants felt that they were not necessarily organized, but because of the limited amount of data they managed, this was not an issue. In fact, frugal participants wanted to minimize their personal data so that they could reduce the time necessary to manage and curate items. They cared about privacy and security, and were open to user-controlled tools that could help them reduce their data, although they often did not know or use any in their daily life. On a more general level, they had a somewhat uneasy relationship with technology, trying to avoid spending time or money on digital devices. For example, it was common that participants said they did not want to pay for storage space and often they used old devices, like flip phones. These participants considered the realm of digital data as secondary and not as important as anything that was physical or disconnected from technology. They talked aspirationally about wanting to live a life free of distractions and intrusions from data and technology.

For example, one participant explained the different approach in curating physical and digital items saying that she did not feel a connection to digital items at all:

"I feel I don't have much of a connection to physical things, but I have even less to digital, right? So, it's like... digital things, I just do not care. Delete, delete, delete. With physical, I would stop and think before I put [them] in the garbage. (P1, Study 2)

6.4.6 Temporal aspects of data curation

One common aspect across behavioral styles was the ongoing nature of curation as a long-term practice. We found that participants curated their data at different moments and with different strategies or approaches for *decluttering* data (that we detail in Chapter 5): *routinely* if they did it on a regular basis, without urgency; *serendipitously*, when they happened to encounter data while in the middle of other tasks; and *urgently*, when they were triggered by events like storage space on their devices running out. Different behavioral styles relied on different strategies more or less often, as we show above.

User strategies often involved looking at time as a dimension of curation, with participants showing both a retrospective and a prospective attitude to data. In a *retrospective* attitude, they looked back and evaluated items based on their use and other subjective criteria. In a *prospective* attitude, instead, they looked forward, trying to predict their future data needs. As one participant explained: "*I used to have a lot* [of yoga albums on my phone] *but now only keep six.* [...] *I had to go through and think whether I would use it in the near future or not. So I only kept six. I picked the best and I got rid of the rest.*" (P4, Study 2)

Over time, participants' approaches often evolved, influenced by life changes and transitions. Participants expressed a need to use curation as a way to evolve and redefine their identity. This process involved reflecting on the past and anticipating the future. It meant re-considering values and re-negotiating attachments. For example, one participant described her approach to organization and curation as "phasic," explaining that her data practices changed over time:

"I am not a person who keeps organizing all of what is getting stored, there isn't the time and attention. Long back I had folders for my professional [things], work. I don't organize that frequently now. Per need, I do. If I am working on a project, or moving, any documents I collect, I would place them in one folder so that I have ease of access, so it's more of a phasic organization. [...] Taking time to look at the pictures: we tend to forget, and that's how the world seems to be moving, click, click, click. Unless you're posting somewhere. There was a time I used to post on Facebook, which I stopped, for several years. It doesn't seem that I need to, I am fine. That also stopped me from accessing my photos and organizing them." (P17, Study 4)

6.5 Evolution of the behavioral styles descriptions

In this section, we reflect on how the behavioral styles evolved over time across the four studies and expand on our process. The analysis process was highly iterative and took several months (in parallel with analysis focused on answering more specific research questions in each study).

After Study 1, we wanted to expand the focus of our investigation and unpack the spectrum of "hoarding" and "minimalist" behaviors. In particular, we decided to tease out the individual variation within "minimalist" tendencies by taking a deeper look at a smaller set of participants. This approach helped us make our initial model of tendencies more complete. But it was still not enough for moving towards design.

Knowing that we wanted to explore possible, alternative design solutions in the space of data curation, we needed an actionable way of referring to the individual differences we saw across participants. We also realized that we needed a way to recruit participants based on their differences, if we were to fully explore the design space. Recruiting a sample with varied approaches to data curation was essential to validate our design process and the idea of building personalized tools for curation. So, after the first two studies, we decided to take a higher-level look at the interviews we had collected and develop a set of user "archetypes," behavioral patterns that could summarize individual behaviors, inform design decisions and help evaluating them. We chose to develop "archetypes" inspired by recent industry work on user modeling [20]. We later renamed them as "user types" and then as "behavioral styles" to avoid an over-simplistic interpretation of these patterns as ideal, fixed categories of behaviors. From the beginning, we avoided thinking of the behavioral styles as personas, as we discuss later.

6.5.1 Initial, preliminary version

The initial work on the behavioral styles took place after Study 2 and a few months before running Study 3. In this first phase, we analysed the 30 interviews from Study 1 and 2 using the process described in Section 6.3.5. The initial set of behavioral styles (at the time still labelled "archetypes") consisted of six instead of five (what would become the "casual" style consisted of two separate but related styles, "inbetweeners" and "untroubled"). Some of the styles also used different labels. For example, "disciplined" instead of "purger".

6.5.2 A first, stable version

Before running Study 3, we used the initial version of the behavioral styles to discuss the framework within the team and our research lab. We asked the opinion of people who were familiar with our line of work as a way to test how the set of behavioral styles resonated with an external audience. In this phase, we realized that some of the labels were not communicating the key characteristics of each style, and that two of the styles (”inbetweeners" and "untroubled") could be merged into one without losing meaning. We went back to the interviews, looked at our codes, and, using constant comparison, we iterated on the labels and segmentation of the styles, arriving at a first, stable version that was very close to the one we present here.

6.5.3 Summarizing the behavioral styles for recruitment

With the five behavioral styles largely stabilized, we used a short description of them in a screening survey to recruit a varied sample of participants in Study 3 (and later Study 4). Respondents could choose one of the five options below, each corresponding to one behavioral style. We used the same set of options in both studies. Note that respondents did not see any label, only the descriptions. The short descriptions capture only the key differentiating aspects of each behavioral style, balancing the need for nuance, with the need for simplicity that a screening survey requires:

> *Casual*: I keep some data, delete some, and generally have a pretty relaxed approach to organization. I don't think or worry much about data management.
>
> *Overwhelmed*: I keep most of my data, but sometimes feel overwhelmed by it. I am not as organized as I probably should be because it is hard or I don't have time to do it. I would like to get rid of some data.
>
> *Collector*: I keep lots of data, I am generally organized, and I am happy about my approach.
>
> *Purger*: I only keep necessary data. I am organized and regularly spend time deleting and managing data.
>
> *Frugal*: I do not have a lot of data to begin with and I try to avoid spending a lot of time with technology.

As we have mentioned before, we do not see a strict one-to-one correspondence between a behavioral style and a person because these patterns are dynamic and contextual: behavioral styles can vary based on data types and over time. However, we asked participants to choose only one option to simplify the survey and also ensure stronger individual variation in the sample.

6.5.4 Validating the behavioral styles with additional studies

With the behavioral styles summarized in short descriptions, we used Study 3 and 4 as a way to validate our first stable version and also enrich it (as we detail below). In both of the last two studies, we had to check the match of participants to the option they picked in the screening survey. We did this by asking them the same interview questions from Study 1 and 2 that we had used to generate the behavioral styles in the first place. Carrying overlapping interview questions across different studies allowed us to check our assumptions with new data and slowly lead to convergence.

In general, most participants in Study 3 and 4 matched what they reported as their main approach when filling out the screening survey. But, as expected, they often had exceptions to their general approach. It was common for participants to display and discuss patterns of behaviors linked to more than one behavioral style, based on different contexts (e.g., work life vs. personal life) and data types (e.g., photos vs. messages). In most cases, participants closely matched one or two behavioral styles. In a few cases, they matched three.

In a few cases the self-reported approach of participants in Study 3 and 4 did not match what they told us and showed us during the study sessions. Often, it was a matter of having a very diversified approach for many different types of data, as we explain above, with the option chosen in the survey not fully capturing this nuance and diversity of data types (as we expected from having a simplified description of the behavioral styles and a single-choice question). In a couple of cases we had participants reporting being "collectors," for example, but when talking to them and seeing their data practices, we felt they were closer to being "overwhelmed," because they used similar expressions and showed similar patterns to other participants we grouped in the "overwhelmed" behavioral style. With the exception of this minority of cases, our screening question seemed a reasonable approximation of actual behaviors but we later note how it should be seen as a tool that needs checking, rather than a definitive match, especially because different data types often call for a different approach. There might also be individual differences in the degree of reflexivity and confidence that different people have around data curation, and out of hundreds of respondents, a few might have answered the behavioral style question, which appeared last on the survey, without paying close attention to all the options (this is always a possibility in surveys).

Another result of using the behavioral styles as a screening tool for recruitment is that both in Study 3 and 4 we managed to recruit only a few participants who displayed a frugal approach. However, we expected this result given our online recruitment process: we did not expect a large number of respondents who largely take a frugal approach to frequently check online recruitment advertisements. Further, the frugal behavioral style is based on a smaller number of participants. Regardless, we felt it was different enough to be part of the analysis.

6.5.5 Connecting the behavioral styles to design research

In Studies 3 and 4 we also used the behavioral styles to drive our design process and, at the same time, we leveraged the resulting design work to enrich the behavioral styles. A key goal of our work across the studies is to provide support for different patterns of behavior in personal data curation. The behavioral styles were essential for defining and scoping these patterns, and then understanding how to inform design decisions. In Study 3 (Chapter 7) we explore different dimensions of curation based on previous work, including automation and system aggressiveness. When exploring how to make these dimensions into concrete design concepts we often referred back to the behavioral styles. For example, the collector behavioral style informed *Patina*, one of the concepts we explore, a visualization of temporal aspects of data (e.g., frequency of use, age). We imagine that a similar concept might play well with users who tend to keep and manage large quantities of data over time. Similarly, we have concepts (*Temporary Folder and App*, *Future Filters*) where data is automatically deleted in the future that were inspired by prospective curation decisions and the need to "purge" items that some behavioral styles referred to.

In Study 4 (Chapter 8), instead, we bring together insights from previous studies into a cohesive prototype system for curating data. Here, we used high-level goals from across the behavioral styles to create a set of filters for sorting through personal data: wanting to find data to purge, wanting to improve organization, wanting to protect sensitive information, and wanting to avoid any potential data loss. We also used different general attitudes about automation to design two different interfaces for achieving the same goals: one focused on providing information but leaving the ultimate decisions about what to do with users, and one focused on automating the data curation process with recommendations. We detail these diverging attitudes in Chapter 7. Drawing on the behavioral styles, we outline a series of use scenarios for evaluating the prototype and we develop additional insights about data curation, that we detail in Chapter 8.

We used the design work from these last two studies to elicit reactions from participants and add more dimensions to the behavioral styles, considering aspects of personal data curation that we had paid less attention to before. For example, in

Study 2 we started to look at how data management tools play a role in supporting (or hindering) data curation practices. We continued to look at tools and different attitudes towards the role of technology in Study 3 and 4. In Study 3 we enriched the behavioral styles descriptions with attitudes about automation, detailing opposing stances toward automated curation tools. In Study 4, instead, we explored aspects of curation related to privacy and security, aspects of curation that we highlight at different points in the behavioral styles descriptions. The shift in methods, from exploratory and contextual interviews (Study 1 and 2), to research through design (Study 3 and 4) allowed us to look at the behavioral styles from different perspectives. We were able to build a deeper understanding of curation leveraging both user narratives [288] (Study 1 and 2) and direct exposure to possible design solutions (Study 3 and 4).

6.5.6 Final version

Between Study 3 and 4 the behavioral styles were largely stable. Several participants from later studies closely matched answers from earlier studies. At this point, we felt the work had reached convergence and the behavioral styles were mature enough. The version of the behavioral styles that we presented earlier is the result of this overarching, iterative process.

6.6 Discussion

6.6.1 How the behavioral styles help us understand data curation

We started the chapter by asking how to make individual differences in personal data curation actionable. The behavioral styles we described fulfill that role as a resource for research and design. But the idea of categorizing people and assigning them labels can be controversial. Categories can hide implicit judgements about behaviors and risk placing people into a singular box that they may not solely fit within. Additionally, due to the subjective nature of data curation, implicit judgements and categorizations are almost impossible to escape. Participants often described their own respective approach by comparing themselves to other people: "*I know people who keep things for years and never go back. I'm not that kind of* *person*" (P15, Study 4), "*Unlike many people I use my phone with extreme caution*" (P4, Study 4). Such statements show how participants compared and judged others to rationalise and communicate their own approach. Personal data curation helped participants define themselves against other people [139].

Both Study 1 (Chapter 4) and prior work show that curating data helps in identity building [66, 68, 140, 149] and that people hold strong moral views about the best way to manage their data [288]. Study 1 introduced the idea of "hoarding" and "minimalism" as two opposite ways of thinking about data and identity. However, this process is contextual and the labels are grounded in participants' words. The behavioral styles expand that model, with different styles matching one or the other extreme: the Overwhelmed and Collector behavioral styles tease out variation within the tendency of "hoarding," the Frugal and Purger behavioral styles tease out "minimalism," and the Casual behavioral style represents a bridge between the two sides of the spectrum (Figure 6.3). They tell us what actions people take to fulfill the core function of identity building at the heart of data curation and show us how this process takes place over time. These insights resonate with theories from sociology, consumer behavior, and social psychology that detail how identity building is a process that shifts and evolves over time through social interaction [17, 65, 194]. People imbue tools and objects around them, both physical and digital, with symbolic meaning [17, 18]. Then, these objects and concepts, with their symbolic meaning, play an important role in their evolving sense of self and in relationships with other people. When we connect these ideas to the way participants discussed their approach as evolving (or being stuck) in time, we can see the behavioral styles as a dynamic categorization. The styles are not a set of boxes. Instead, they are a set of patterns that can shift, mix, and evolve, just like people do. Below, we discuss the implications for design and research of conceptualizing behavioral styles as we do.

6.6.2 Behavioral styles as a resource for design and research

A common argument against the type of user modeling we present in this chapter is that people are complex and no behavioral style will ever fully capture their nuanced experience. That said, we argue that the behavioral styles we built can be a

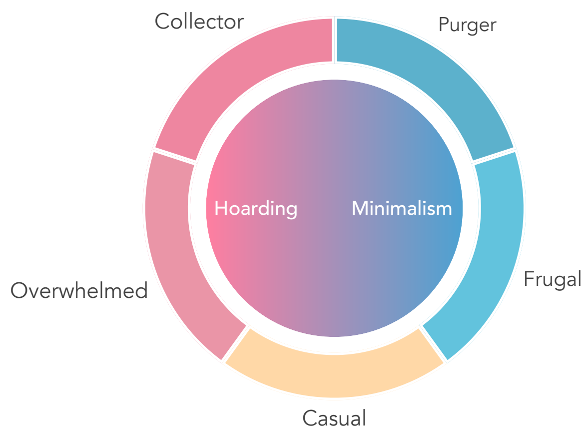


Figure 6.3: A visual representation of the five behavioral styles as an expansion of more general data curation tendencies: the Overwhelmed and Collector align with "hoarding" tendencies, the Frugal and Purger with "minimalist" tendencies, and the Casual style bridges the two sides.

useful resource to drive design decisions, inform research recruiting, and help with qualitative analysis. In our work across studies, they were an essential resource that we turned to when we did not have direct access to participants. They were relatively inexpensive and they helped us ground our process. But it was essential to see them as a tool: something you pick up when you need to, but that you put back when its job is done. To better explain our framing of behavioral styles as a design and research tool, we compare them to personas.

Why did we not use personas? We could have. Personas are a common method in user-centred design, promising to drive decisions and build empathy. We could have created typical personas like "Harry, the Happy Hoarder," or "Mary, the Minimalist." But we did not. We we decided to avoid thinking of these behavioral patterns as personas because of personas often focus on demographics and assume that individual behaviors are static. While recent work has looked at integrating complex identities into personas [180] or making them more participatory [204], the arguments against them are plenty [179]. Personas can be abstract and impersonal [189], they can reiterate gender stereotypes [131] or power relationships [70], they can be removed from empirical data [88], and they are difficult to validate [53].

Compared to personas, behavioral styles helped us focus on contextual patterns of behaviors. We were not trying to help and satisfy a single imaginary person. Instead, we were trying to support behaviors grounded in specific scenarios we observed. This approach influenced our design outcomes and led us to solutions that opened a design space. For example, we could have created a tool that would categorize one person as a typical user or persona based on demographics and behaviors, and then personalize the interface according to fixed, preset criteria. We did not, because we saw behavioral styles as contextual patterns of behaviors that only represent a starting point. Similarly, when we used the behavioral styles as a recruiting tool, we saw them as a starting point for reaching variety in the sample. But they were not an end in themselves. After recruiting we still had to assess the fit of each participant to one or more behavioral style, and understand what were their exceptions to general patterns of behaviors. This approach made it possible to generate unexpected design opportunities that build on the intersection of different behaviors and contexts.

Our conception of behavioral styles as a representation of dynamic behaviors and a generative resource calls into question common assumptions about user modeling. Often as designers and researchers we turn to modeling methods such as personas in search of a prescriptive and predictive model that can help design by limiting its focus. Instead, we propose to look at behavioral styles and similar modeling efforts as a generative and descriptive resource. The goal is not to reduce design options, but instead opening new opportunities. In the next section, we outline some opportunities based on our work, expanding on possible design efforts for data curation tools and future research investigations around personal data.

6.7 Opportunities for design and research

6.7.1 Prioritizing user support

A first approach in design could be to prioritize different patterns of behaviors, based on the level of support they need. For example, we see "overwhelmed" users as the ones needing most support. Participants who felt overwhelmed reported feeling stuck and resigned, but wanting to change their approach. In a sense, they felt left behind by technology. Exploring new tools targeted at this style of behavior is a key opportunity. Recently, tech companies have introduced of tools that can proactively find and recommend items to declutter from devices or cloud platforms (e.g., Google Photos, Files). This is a reasonable approach, but one that can undermine agency (as we show in Chapter 7). An alternative would be to inform users of different curation practices and help them learn how to do things differently using their current tools. This approach could take the form of assistive learning tools integrated in existing user interfaces. Automated suggestions could show users possible alternative ways of organizing or curating data based on different patterns of behaviors. This is exactly how the behavioral styles can help generate ideas and integrate patterns of behaviors into design work. For example, the system could show to "overwhelmed" users how a user with a "purger" approach would re-organize and declutter their data. These tools could then explain the rationale behind the suggestions, and let users try and explore the proposed approaches before committing to them. A different approach could be an illustrative, informational guide, similar to related efforts about promoting awareness of backups [313] and privacy [200], that would illustrate different approaches and strategies found in PIM studies as a form of direct knowledge-transfer from empirical research to users.

6.7.2 Making curation feel more engaging

Other behavioral styles point to opportunities for making data curation more fun and engaging. This approach could be compelling for some casual users while also matching the need to purge data of other behavioral styles. Some ideas include gamifying the curation process [320]; imagining new ways for repurposing and "recycling" old data (for example, transforming one data type into a different one: texts becoming haikus is a recent example [258]); or, allowing users to donate data, willingly, to projects that have some collective benefit (e.g., research projects, a historical archive), with proper mechanisms for anonymization. This last design direction builds on investigations that have explored a similar idea in a different context (e.g., donating personal data after dying) [151]. The emphasis would be on creating engaging experiences that offer benefits beyond optimizing space or organization.

6.7.3 Exploring the role of curation for memory

Yet another opportunity is to better explore the role of personal data for memory and how to make devices or cloud platforms feel like lived-in spaces rather than static repositories. Features like "Remember This Day" on Facebook and "Rediscover this day" on Google Photos seem to address this need [243]. But they also cause tensions around the evolving nature of personal identity [252]. A more subtle approach could focus on segmenting and resurfacing data based on temporal dimensions [57, 209]: for example, grouping data based on calendar-based periods (e.g., songs I listen to on Tuesdays) or phases of the day (e.g., data I interact with on rainy nights). Services like Spotify provide similar personalized re-packaging of music playlists [262], and we believe that similar ideas could be especially relevant for users who take a collector approach. There is still a lot to explore about how to best integrate these ideas in the design of data management tools that include broader types of data [210].

6.7.4 Investigating new ways of capturing user variation

As we mentioned, we decided to condense the behavioral styles in a single question to make the screening survey easier and quicker to answer: we had additional questions to include and we wanted to keep the survey as short as possible. However, investigating new ways of including the behavioral styles in the recruiting process is a potential avenue to explore in future research. Studies in psychology that focus on capturing individual traits often use scales with several questions. A similar approach could focus on wording the different behavioral styles as dimensions that participants can measure themselves against. This approach could make it possible to better capture the composite nature of individual behaviors. Still, an added complexity when looking at data curation is the way different approaches and different types of data intersect. Building scales with multiple items for multiple data types might increase precision but also add complexity for respondents. Another opportunity would be to capture the behavioral styles in a visual way (for example, showing typical device and cloud platforms configurations for each style), letting users choose the one most similar to their own approach. A visual representation of different behavioral styles and strategies could also act as a probing tool during user interviews, helping participants better articulate their own curation strategies by comparing them against established patterns.

6.8 Limitations

We do not see our work as globally valid or as an absolute truth about people's behaviors. Instead, similar to what we stated in Chapter 4, our analysis reflects the social, technological, and cultural context of the participants we talked to: predominantly working-class folks, all living in a major Western country. These parameters should inform the *transferability* of our results [110, 165, 280] to another context. Future studies can complement our work by looking at similar research questions in a different context, to unpack differences and overlaps across cultures and generations. We also encourage future studies to take a different epistemological approach for investigating individual differences in personal data curation. We took a constructivist, interpretive, qualitative approach [110, 165] grounding our analysis in participants' experiences and narratives [288]. Future research can build on our work by collecting quantitative measures of user actions (e.g., number of times a user moves, deletes, or re-accesses data on a given period) and explore emerging patterns of behaviors. Some previous user modeling studies, for example, identify user clusters using surveys and correlational analysis [223, 279]. A similar line of work could focus on exploring the prevalence of different behavioral patterns in a broader population using quantitative data. Once again, we note that we did not see a rigid one-to-one correspondence between a participant and a behavioral style and we caution against a similar interpretation of our work, but we encourage future studies to elaborate on this conclusion by using different methodologies. Finally, an additional way to build upon and expand our contributions is to take a longitudinal approach and better capture temporal aspects of data over long periods of time (e.g., months or years.)

6.9 Summary and conclusion

Personal data curation is an underexplored topic with rich implications for the design of data management tools. As society moves towards a world where digital data will become even more pervasive, it is essential to understand individual user needs and accommodate the different ways people manage their personal data. We have proposed a set of individual user behaviors, describing five behavioral styles and their approach to curation. The behavioral styles represent an actionable resource for design and research, opening the way for new products and future investigations in this domain. Leveraging our work, future research can tailor design efforts to different users and explore innovative ways of thinking about personal data.

In the following chapters, we build upon the behavioral styles to explore some of the design directions we outline. In particular, in Chapter 7 we use the temporal dimensions of personal data curation and different behavioral styles to inform the design dimensions and concepts we explore. Then, in Chapter 8 we focus on prioritizing and personalizing user support for data curation.

Chapter 7

Study 3

Exploring a Design Space for Selecting Personal Data

With this chapter[1], we enter the second half of the dissertation, where design work takes the stage. Here, we build upon the insights from previous studies and explore how to support the selection of personal data. We define *selection* as the practical process of choosing what to keep or discard. As in previous chapters, the actions of selecting what to keep or discard is a subset of the *keeping* stage in personal data curation.

7.1 Introduction

So far, we have looked at personal data curation and the different ways people approach the process, driven by general tendencies (Chapter 4), using several decluttering strategies (Chapter 5) and reflecting different styles of behaviors (Chapter 6).

In this chapter, we want to look specifically at how technology can support *data selection*, the process of intentionally deciding what personal data to keep or discard. We have argued before how selecting what data to keep or not is a

[1] Originally published in Vitale, Odom, and McGrenere. (2019) *Keeping and Discarding Personal Data: Exploring a Design Space*. Proceedings of the 2019 Conference on Designing Interactive Systems (DIS '19) [291]

necessary step for "emotionally viable" archival systems: choosing what to keep and what to let go is important so that you can derive value from your personal data over time by keeping things that matter to you [182]. We have also see how it has become nearly impossible to decide what to keep and discard [24] due to the growing amount of personal data.

How can technology support this selection process? Recent studies on the value of digital data [111] and its longitudinal management [150] point to a growing need for tools that can support keeping and discarding decisions. But we know from previous chapters and previous studies [150] that people show strong individual differences in their practices, so it is unlikely that a single solution could satisfy all users.

With these premises in mind, we ask: How can technologies be better designed to support people's decisions around what data to keep or discard? In particular, what different design approaches might be viable to different users and in different situations?

Using a Research through Design approach [317], we created five design concepts as a way to probe people's reactions, attitudes, and perceptions on the role of technology in supporting personal data selection practices. The concepts intentionally emphasize different design dimensions stemming from related work in PIM. The five concepts are: *Patina*, a visualization of temporal aspects of data (e.g., age or number of interactions), *Data Recommender* (a recommender system that suggests data to take care of using machine learning), *Temporary Folder* (a folder with an expiration date), *Temporary App* (a mobile application with an expiration date), and *Future Filters* (a mobile application to create advanced filters for deciding what to do with data in advance). For each concept, we created a short video sketch [316] as a prototype that primarily illustrates how it works. Then, we conducted one-on-one interview sessions with 16 participants with varied data management approaches. The interview sessions touched on the potential benefits and drawbacks of the concepts, with a range of reactions. We identified contrasting attitudes towards the systems we presented, with the tension between automation and control informing the need for context-based solutions. Drawing on the interview analysis, we critically reflect on our results and outline future design directions to further open the design space.

In this chapter we make three key contributions. First, we outline four design dimensions (*selection regime*, *automation*, *aggressiveness*, and *temporality*) to define and broaden the design space around keeping and discarding decisions. These can be used as a generative resource for creating new solutions. Second, we offer five alternative design concepts that we used in an elicitation study with a diverse sample to probe and explore the space, showing where people's key boundaries around control and automation lie. Third, we discuss future design directions for supporting keeping and discarding decisions focusing on personalization, automation, defining new actions, and targeting data privacy.

7.2 Related work

Below, we review design efforts around personal data management from prior research projects. We use this review to outline a set of design dimensions and approaches to probe on. For more general insights about user practices and challenges in selecting data, we refer to Chapter 2.

7.2.1 Existing and proposed design approaches

Augmenting data management interfaces

Two common user interface paradigms to manage data are: 1) *document-centric*, with the most dominant desktop metaphor of files and folders (common on personal computers and cloud storage platforms), and 2) *application-centric*, with the application acting as a bundle for data (common on mobile devices and social media platforms) [13, 301]. Both paradigms have benefits and drawbacks. The file systems community, for example, has criticized the desktop metaphor organized around folders for being too rigid and inadequate as the amount of data grows [255]. But despite their faults, folders still dominate management platforms because they provide valuable functions: they help people control, organize, and structure their work [301]. Enhancing them, rather than replacing them, might be the best design approach [301].

Several projects propose augmentations or alternatives to folders (Table 7.1), using for example metadata [300] or annotations [294]. Other projects choose an alternative activity-centric approach [13], exploring "time-ordered streams" of documents [90], flexible desktop organizations [79, 295], or new metaphors based on places, time, and data provenance [167]. While exploring radical alternatives can help push forward design, our concepts largely focus on augmenting current interfaces so that participants can better relate them to their own experience.

Design approach	**Projects**
Document or folder augmentation	BIGFile [168], Facet Folders [300], File Biography [167], Finder Highlights [93], GrayArea [29], Old'n Gray [31], Project Planner [143], Vanish [97], WikiFolders [294]
Filtering or tagging	DMTR [30], Phlat [69], Stuff I've Seen [82], Tagtivity [221]
Activity-centric	File Biography [167], Giornata [295], Lifestreams [90], Presto [79], Tagtivity [221], Task Aware Ranking [281]

Table 7.1: Previous research projects categorized based on their general design approach.

Using automation to complement selection practices

Most of the design projects reviewed so far focus on data organization or retrieval, with little attention paid to keeping decisions (see Table 7.2). An exception is work by Bergman *et al.*, with several related projects addressing the keeping stage of the curation cycle: *GrayArea* [29], *DMTR* [30], and *Old'n Gray* [31]. These systems use the "demotion" principle [22], an intermediate action between keeping and deleting (which is the most common discarding action afforded by user interfaces). *Demotion* makes items visually less prominent or hides them in a separate area of the interface. This is a valid compromise between keeping and deleting: unnecessary items do not distract when they are demoted, but they are still there in case they are ever needed. Although *discarding* data can mean more than deleting (e.g.,

Curation stage	Projects
Keeping	DMTR [30], GrayArea [29], Old'nGray [31], Vanish [97]
Organizing	Giornata [295], Lifestreams [90], Presto [79], Project Planner [143], Tagtivity [221], Wiki-Folders [294]
Retrieving	BIGFile [168], DMTR [30], Facet Folders [300], Finder Highlights [93], GrayArea [29], Lifestreams [90], Old'nGray [31], Phlat [69], Presto [79], Project Planner [143], Stuff I've Seen [82], Tagtivity [221], Task Aware Ranking [281]

Table 7.2: Previous research projects categorized based on the curation stage they focus on.

demoting) our concepts focus on deleting as a key action to elicit more powerful reactions from participants and understand where their boundaries lie.

The examples by Bergman *et al.* also highlight two distinct design approaches to solve the "burden of curation" [140]: in GrayArea, users rely on direct manipulation to demote items by dragging them into a separate area of a folder, where in DMTR and Old'n Gray the process is automatic. The tension between automation and user-control is at the centre of many investigations in HCI [91, 102, 136, 237]. Within PIM, the discussion around automating user management strategies is key [286] and goes back to early studies about email [19, 304]. Some examples of automation or semi-automation focus on selecting photos [164, 208] and audio [195] or on the process of passing down digital data [112]. Jones [142] discusses automatically archiving information that is no longer useful, while *Vanish* [97] introduces the idea of self-destructing data. Bergman *et al.* [26], among others, argue for finding a balance between automation and user control in PIM interfaces. Yet, this is a question rarely explored in the specific context of keeping decisions. Understanding which "curating" actions can be automated and which should not is an ongoing open question in PIM research [144]. Our work uses the tension between manual actions and automation as a key design dimension to explore potential new directions.

Using metadata to build awareness of digital items

Most design efforts discussed so far focus on improving data management tasks. A different strand of design work by Odom and colleagues, instead, offers a counter-perspective, and focuses on reflection, reminiscence, and enjoyment [212]. This approach uses metadata for rediscovering kept data through everyday objects [209, 218]. Sas *et al.* [247, 249] also use a similar, reflective approach by proposing "rituals" for letting go of sentimental digital items. We use this work as inspiration to add possible design choices and incorporate an open-ended, reflective dimension in some of our concepts. While prior work focuses on tangible interactions, we explore using metadata in graphical user interfaces to resurface old digital items or build awareness of their accumulation over time.

Exploring the potential of prospective decisions

Finally, we narrow the focus to email, with work by Gwizdka [114–116] providing additional inspiration. In studying and designing email management tools, Gwizdka introduces the idea of *prospective* information to support task management by anticipating future needs. Today, several email management tools apply a similar concept through reminders and snoozing functions. Empirical work on more general data selection decisions points to temporal dimensions that consider both the past and the future [140]. For example, Kim [151] mentions one participant having a folder called "to delete," while Brewer *et al.* [42] discuss prospective memory for digital reminders. Our own work in Chapter 6 outlines *retrospective* and *prospective* dimensions of data curation. Thus, we see the challenge of anticipating future needs with prospective decisions as a major opportunity to expand the design space. While most data management tools are retrospective, there might be space for prospective decisions and we use some of our concepts to explore this area.

7.3 Methodology

7.3.1 Research approach and design dimensions

Our review of related work shows how designing to support keeping and discarding decisions is a largely under-explored territory, with different potential directions to follow. This multitude of possibilities makes a design-led exploration ideal. Thus, in our work, we took a Research through Design approach [317]. Our inquiry can be seen as parallel to that of Gulotta *et al.*, [112] who use of a similar approach to investigate the space around data curation, legacy, and memory.

We started by clustering and mapping insights from prior work and our own empirical work described in previous chapters into four key design dimensions to probe: *selection regime*, *automation*, *aggressiveness*, *temporality*. These four design dimensions synthesize both previous related work on digital data and our own work in previous chapters. Below, we describe the four dimensions, explaining what we based them on and how we used them to drive the design process.

Selection regime - The first dimension considers possible "selection regimes" [140] that people use when curating data: whether they consider one item at a time or a collection of items all together. This dimension is based both on past related work and our own empirical work. Past related work on curation [140] introduces the idea of *selection regimes*, where people consider items either in groups or individually. In our own work, we saw similar user behaviors in Study 2 (Chapter 5). We created variations along this dimension to encompass both individual items and collective categories, probing on the differences in support needed for both.

Automation - The second dimension focuses on the tension between user-initiated data selection and automated data selection. This dimension is based on past related work on data management: earlier in this chapter (Section 7.2.1), we discussed how past work on PIM explores different approaches to automation, arguing for a balance between automation and user control. Thus, we see automation as a key dimension to explore given the increased potential for automatic data management tools, thanks to machine learning and artificial intelligence. In our design work, we contrasted concepts that took automation to its extremes with others that give more emphasis to user influence.

Aggressiveness - The third dimension is about the level of aggressiveness of the system. This dimension is inspired by a large body of past related work on digital data and metadata, covered earlier in this chapter (Section 7.2.1), that argues for a calm, open-ended approach to designing for personal data. We used this dimension to contrast concepts that are more open-ended (i.e., they only inform of data to take care of, whether the user notices it or not, letting them decide what to do and when) with others that are more forceful and push the user to decide whether to keep or discard something.

Temporality - The final dimension represents the temporal user mindset in selecting data: either *retrospective*, looking at items based on past use, or *prospective*, looking at items based on future use. This dimension is based both on past related work and our own work. Earlier in this chapter (Section 7.2.1), we provide examples of past work that points to prospective data management decisions as a promising area to support. In our own work, we also highlight temporal dimensions of data curation (Chapter 5, Chapter 6), highlighting a contrast between retrospective and prospective user attitudes. We used variations along this dimension to probe on the largely underexplored area of prospective decision making to see whether this might be a viable direction compared to the retrospective nature of many traditional management tools.

After defining the design dimensions, we created five concepts that differ along their extremes to explore a design space for data selection. As we already mentioned, our own work informed the design dimensions, but these dimensions do not have a precise one-to-one correspondence to the five Behavioral Styles presented in Chapter 6, because the dimensions synthesize attitudes from across the behavioral styles. Similarly, when we created the design concepts we considered the five behavioral styles and how each of them might relate to a concept, ensuring that we covered a range of attitudes in the design space. Some concepts have a more explicit link to a specific behavioral style (that we point out in their descriptions below), but our goal was not to create or test a rigid one-to-one correspondence, given the contextual and dynamic nature of user behaviors.

For each concept, we created a short video prototype or sketch [316], an illustration of how it works. The videos use a mix of descriptions and user scenarios, depending on what we felt best illustrated each concept. The concepts take inspiration from existing work or systems, as we detail in their respective descriptions. However, to have more control in our elicitation study, we decided to design our own set of video prototypes instead of using existing systems. By creating our own concepts and videos we were able to push the design dimensions in specific directions, often exploring their extremes in new combinations (Figure 7.1). All together, the concepts synthesize a mix of disparate ideas into a cohesive collection, applying existing and proposed design approaches in new contexts. In the videos, we tailored the user scenarios around key research questions we were interested in exploring with participants. The videos, similar to experience prototypes [45], frame the concepts in a way that offer glimpses into possible futures to provoke and open dialogue with participants about perceived benefits and consequences of each design [178]. Our approach is inspired by and shares similarities with prior work on *User Enactments* [214] and *Speed Dating* [72]. These related approaches argue for exploring the potential roles, values, and social boundaries of emerging near-future technology by using more than one design vision. They encourage participants to imagine future interactions and react to them by drawing on their own experiences.

7.4 Design concepts

We now describe the five design concepts, positioning them within the design space and pointing to sources of inspiration.

7.4.1 Patina

The first concept, Patina, is a visualization on top of data in the geometric form of a spiral [2]. It is inspired by a tree's growth circles and symbolizes temporal qualities of data. In the video for Patina, we show two different options for the spiral: on a desktop, it represents the age of folders (Figure 7.2); with a set of music playlists, instead, it represents the number of interactions over a period of time (Figure 7.3).

[2] Patina's video is available in the supplementary materials to the thesis.

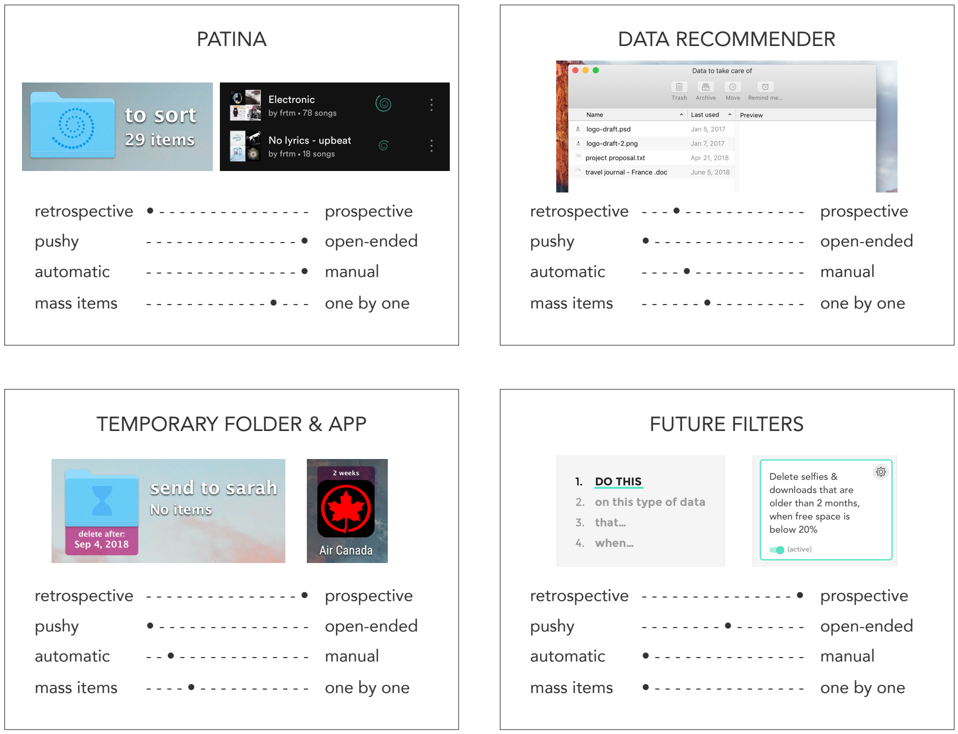


Figure 7.1: An overview of the design concepts and how they map to the four design dimensions we used in our exploration.

Music playlists provide a good contrast to folders because they are "meant to be enjoyed repeatedly and grown over time." [169]

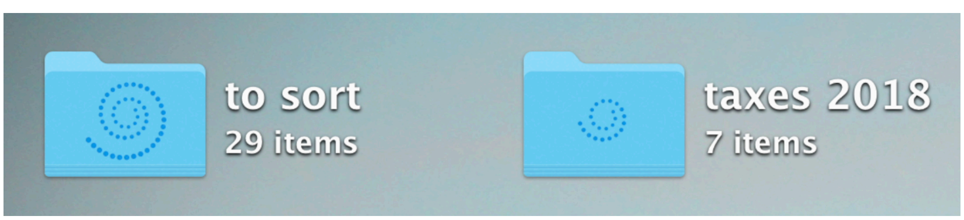


Figure 7.2: Patina showing the age of desktop folders: the bigger the spiral, the older a folder is. (Each dot stands for a set time amount, e.g., a week.)

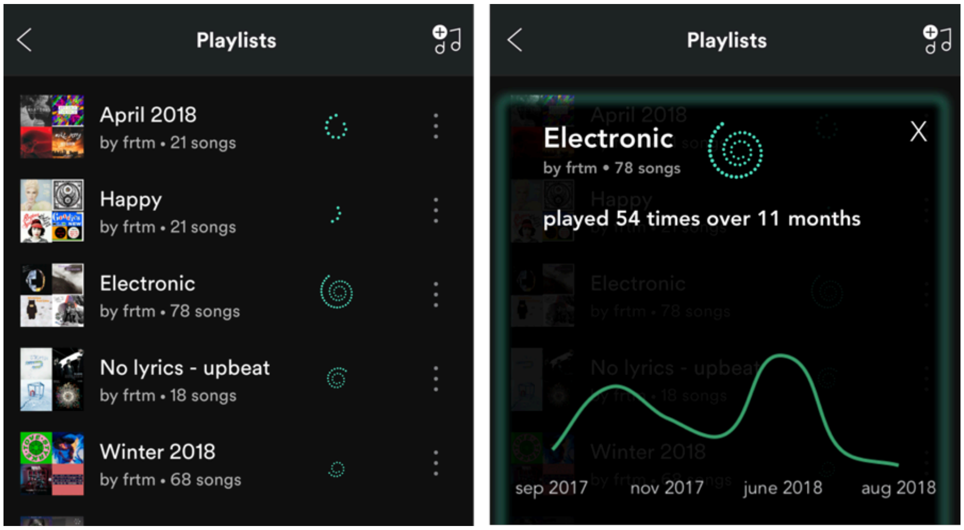


Figure 7.3: Patina showing the frequency of use for music playlists: the bigger the spiral, the more times a user has played the playlist.

Patina's video leaves some aspects as intentionally ambiguous and unexplained (e.g., How is the age of a folder calculated? How is the interaction period determined?). We wanted to encourage user interpretation and discussion. This choice also emphasizes the open-ended nature of Patina in the design space: it invites reflection and builds awareness, but does not suggest any specific action to take on data. We use this concept to probe on the viability of open-ended designs that leave users in charge of initiating any selection action and on what metadata attributes might be useful for doing so.

The idea of a *patina* takes inspiration from prior studies that mention its potential [211] or use it with physical [99, 160, 161] and digital objects [137, 188]. The frequency of use in Patina is inspired by Hurst *et al.* [137] and Matejka *et al.* [188]. The idea of aging, instead, is informed by Giaccardi *et al.*'s work on "traces of use" for daily objects [99]. Our work, however, takes place in a different context and uses a different approach designed with data selection in mind. There are also commercial products with visualizations for used space on hard disks (e.g., Daisy Disk

[3], Disk Inventory X [4]) but these are separate from the data and use only one type of metadata (size). Instead, we tie the visualization to the data and use two types of metadata (creation date for folder and frequency of use for music playlists).

Patina was also informed by the Collector behavioral style. As we mention in Chapter 6, we imagined that this concept might play well with users who tend to keep and manage large quantities of data over time, providing them with an opportunity to revisit and enjoy their collections. We also thought that a visualization such as Patina could engage people who take a casual approach to their data, providing a fresh and unusual perspective to daily data curation. Finally, we saw Patina as a potential way of directing overwhelmed users and users taking a purger or frugal approach to items that they might want to get rid of. A similar visualization could easily match *routine* and *serendipitous* decluttering practices of different users (Chapter 5).

7.4.2 Data Recommender

The second concept, Data Recommender (Figure 7.4), notifies users and provides recommendations on data that might need attention, using metadata such as last access, creation date, or size [5]. Users can decide to trash items, archive them in a central archive, move them in a specific folder, or be reminded of them at another time. Data Recommender will use machine learning to learn from their actions and provide new recommendations. This concept is the closest to existing products. For example, Google Photos [6] provides recommendations on photos to archive, while Files on Android [7] gives recommendations on how to free up space.

Data Recommender is in the middle between human-driven activity and automation, following a *mixed-initiative* approach [46, 134]. Using this concept we want to probe the link between data and context, the viability of different selection actions, and the attributes that make items good candidates for disposal.

When creating Data Recommender, we imagined it might be particularly helpful for supporting *routine* decluttering (Chapter 5) and users who feel overwhelmed

[3] http://daisydiskapp.com
[4] http://www.derlien.com
[5] Data Recommender's video is available in the supplementary materials to the thesis.
[6] https://www.google.com/photos/about/
[7] https://files.google.com

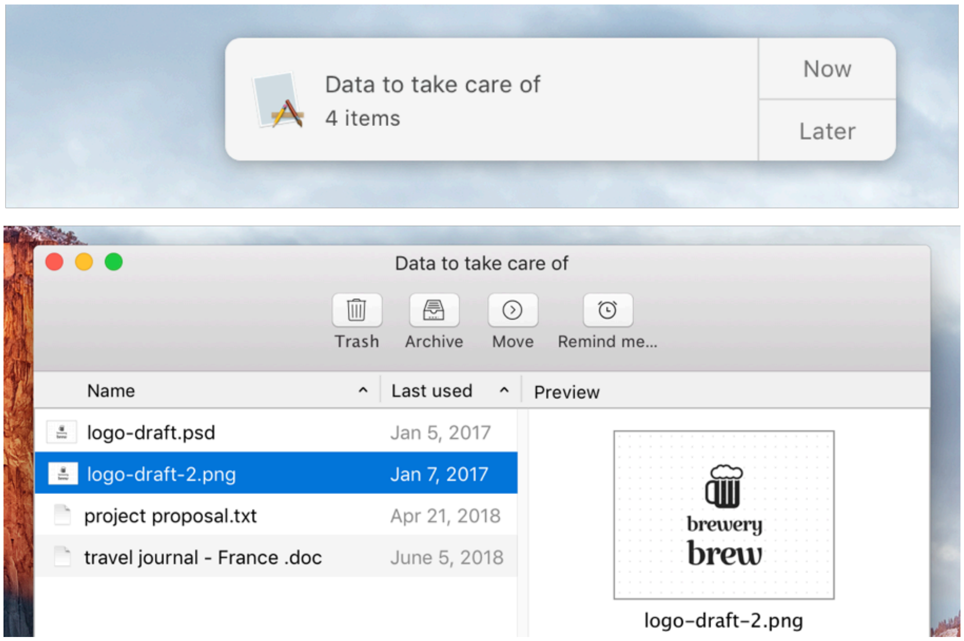


Figure 7.4: Data Recommender notifies users when they have some data to take care of (top) and provides a list of items (bottom): users can choose to trash, archive, move, or be reminded of items again.

(Chapter 6). We also thought it could provide some support to users who take a purger or frugal approach, given their desire to get rid of items. Prior work [215] also points to recommender systems as a good approach to offload the work of selecting items.

7.4.3 Temporary Folder and Temporary App

The next two concepts come as a couple: Temporary Folder and Temporary App. In this case, we created two videos on two different platforms. The first, Temporary Folder, takes place on a desktop computer (Figure 7.5) [8]: it acts as a standard folder, but users can decide to set an expiration date for it. After the expiration date, the folder will be automatically deleted. The second, Temporary App, takes place on a smartphone (Figure 7.6) [9]. In this case, users can install a mobile application

[8] Temporary Folder's video is available in the supplementary materials to the thesis.
[9] Temporary App's video is available in the supplementary materials to the thesis.

temporarily (e.g., for two weeks). At the end of the preset period, the application will be automatically uninstalled.

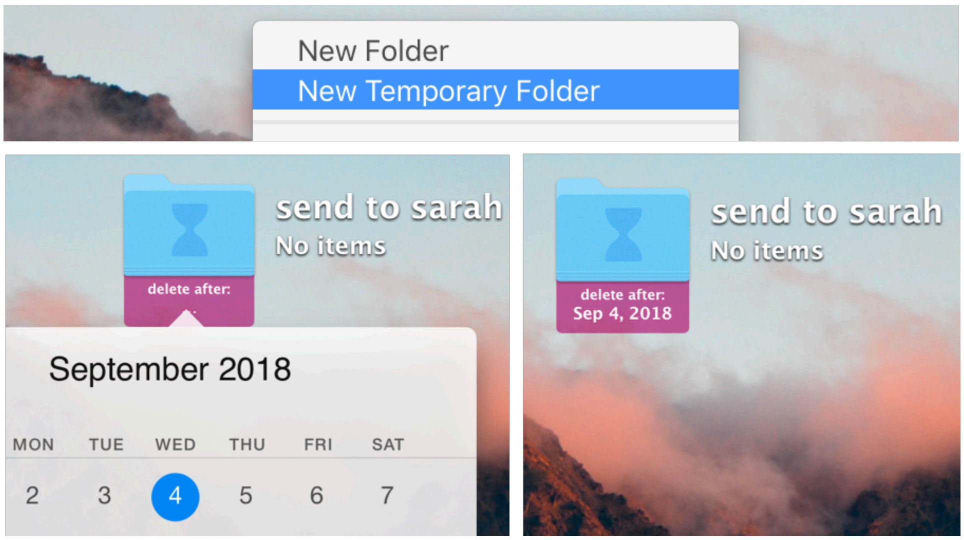


Figure 7.5: When creating a Temporary Folder (top) users can pick an expiry date for it (bottom, left). After the expiry date (bottom, right), the folder is automatically deleted.

These temporary concepts fall within the prospective side of the design space, unlike the previous two. Their radical take on automatic deletion was meant to stimulate discussion among our participants about their perceived social acceptability. The purger and frugal behavioral styles (Chapter 6) were a key influence on these two concepts. We imagined that Temporary Folder & Temporary App could provide support for purging habits and the need to prevent data accumulation in the first place. The preventive nature of these concepts could also help overwhelmed users and could support *routine* decluttering practices (Chapter 5), helping people set prospective routines.

Several commercial products also use automatic deletion in specific contexts. As an example, the messaging application Telegram [10] allows users to set an expiry date for photos, videos, and other files exchanged with contacts. If they do not access them for a set period of time they are removed from the device (they are still

[10] https://telegram.org

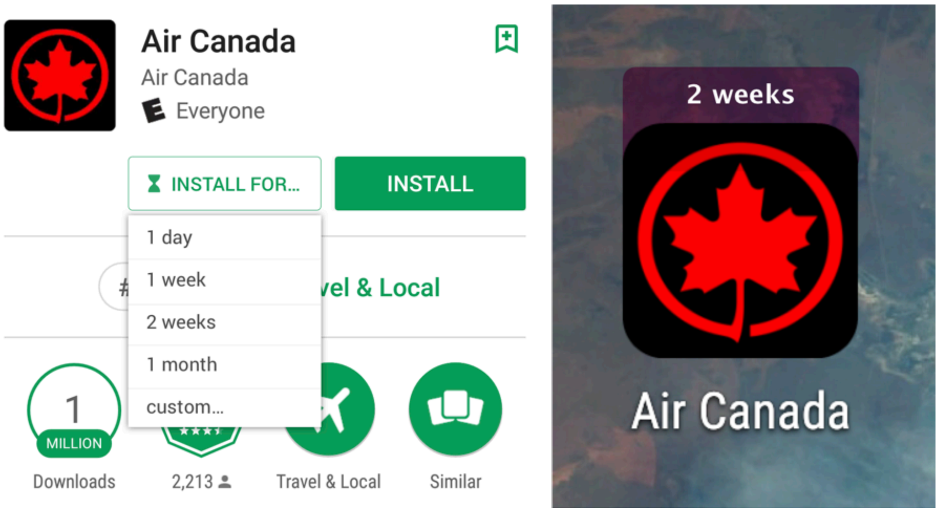


Figure 7.6: When installing a Temporary App users can pick an expiry date for it. After the expiry date, the app is automatically uninstalled.

on Telegram's cloud though, so they are not completely deleted). Snapchat[11] has instead popularized the notion of emphemeral information as a default. The notion of different information lifespans also goes back to an early study of desktop usage that identified three types of information [14]: *ephemeral*, *working*, and *archived*. More recently, Murillo *et al.* [202] explore the potential of data expiration for supporting users' deleting decisions: one participant in their study mentions the idea of an email folder that allows users to set an expiration date [12]. Temporary Folder and Temporary App explore this idea in two specific contexts: desktop files and mobile applications.

7.4.4 Future Filters

The final concept, Future Filters, is a mobile application that allows users to decide what to do with data in the future creating set of rules or filters [13]. For example, "delete selfies and downloads that are older than two months when my free space is

[11] https://www.snapchat.com

[12] The study by Murillo *et al.* was published in 2018, at the same time as we were developing our concepts. We were not aware of it when we designed the Temporary Folder concept or when writing the paper that this chapter is based on. It is nice to see the same idea in two independent studies.

[13] Future Filters' video is available in the supplementary materials to the thesis.

below 20%," (Figure 7.7) or "archive shared documents not looked at in 2 years," and so on. Filters use a set of actions (e.g., delete, move, archive, remind me), criteria (size, use, number of copies, source of data, copied on the cloud, etc.), and triggers (a new update available, free disk space is below a certain amount, etc.).

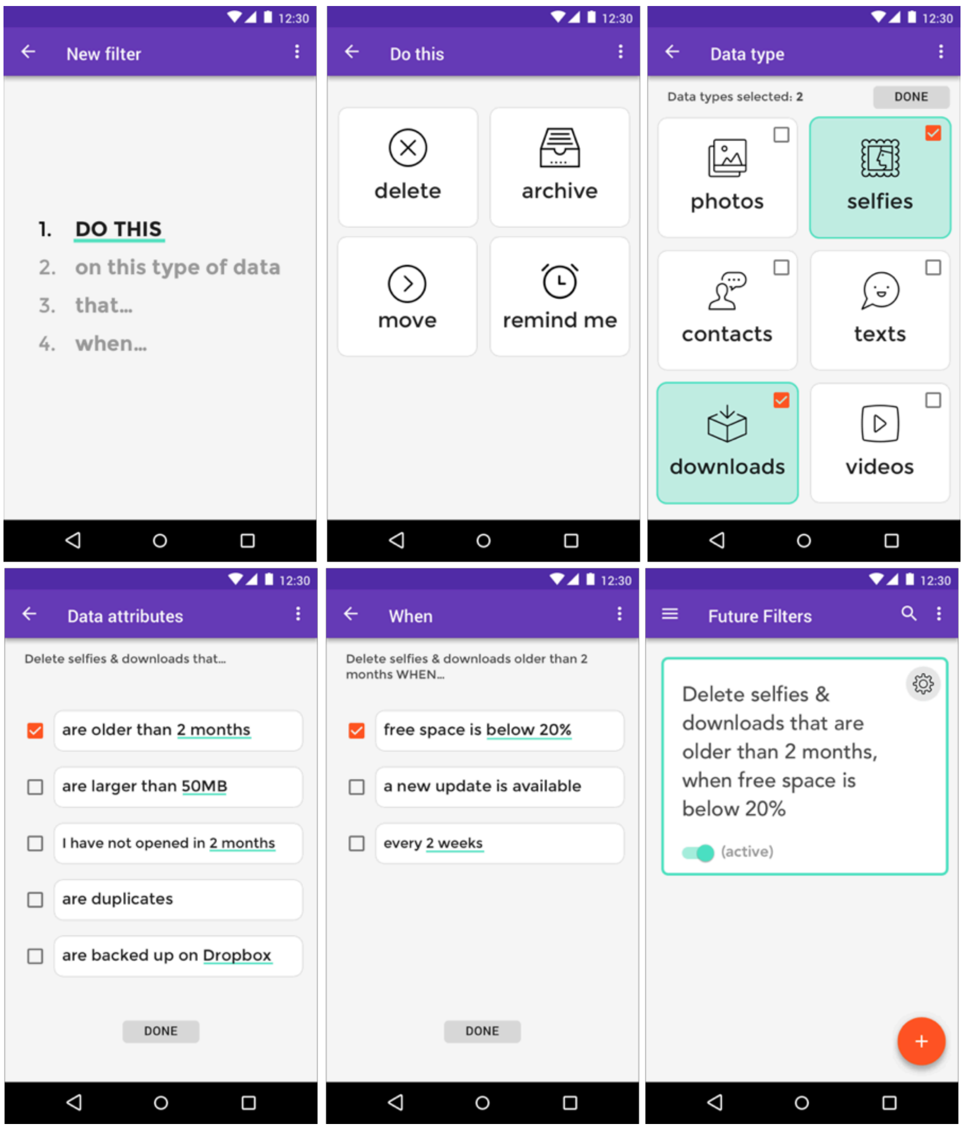


Figure 7.7: Future Filters is a mobile application that lets users create data filters based on actions, data types, data attributes, and triggers.

This final concept has a strong emphasis on prospective decisions and mass processing of items, with a certain degree of automation. The disciplined attitudes to data curation that characterized both the collector and purger behavioral styles influenced Future Filters. We imagined a similar approach could support the need for feeling "on top of things" that participants reported. We also thought this last concept could provide support for *triggered* decluttering (Chapter 5), helping people identify categories of items quickly.

Future Filters also takes direct inspiration from *If This Then That* [14], a platform to create cross-application rules based on triggers, and other products or features that use automatic filters (e.g., File Juggler [15], Hazel [16], Gemini [17], or email filters in Gmail.

We use Future Filters to further explore the viability of prospective decisions, probing on what actions might be more acceptable when considering automation.

7.5 Elicitation study

We used our set of design concepts in an elicitation study with 16 participants. We showed them the videos of the concepts during one-on-one interview sessions that also touched on their general data management practices.

7.5.1 Recruitment and participants

We used purposive sampling to recruit a diverse sample of participants. We advertised the study on a university listing and on Craigslist in Vancouver, Canada. We used a screening questionnaire (see Appendix D) to select participants based on their age, occupation, technical familiarity, and general approach to data curation. We used brief descriptions of the behavioral styles (Chapter 6 in the screening questionnaire as a closed-ended question (see Appendix D).

We received 177 responses to the screening questionnaire. We contacted 36 respondents and 16 agreed to take part in the study. We stopped recruiting when we reached a diverse set of participants and reactions. Nine participants self-identified

[14] https://ifttt.com
[15] https://www.filejuggler.com
[16] https://www.noodlesoft.com
[17] https://macpaw.com/gemini

as female, six as male, one as gender non-conforming. They were aged 23-71 (average: 36). Occupations included administrative assistant, engineer, HR specialist, journalist, photographer. In terms of general approach to data curation, four largely had a collector approach, five felt overwhelmed, three tended to purge items, four tended towards a casual approach. We note that this was the general self-reported approach, but during the interviews participants elaborated on their approach, displaying some differences among data types and more nuanced behaviors. A few participants also displayed a frugal approach in addition to the main approach they had selected.

7.5.2 Procedure

The study sessions consisted of: 1) a short introductory interview on general data management practices, 2) a main elicitation section going over each of the design concepts, 3) a final, longer semi-structured interview discussing and comparing all the design concepts and the ideas behind them. In the introductory interview, we asked participants to discuss how they organized and selected their data over time and on different platforms or devices, asking them to show us examples where possible. Then, for each concept, we first gave a short introduction and then showed the video. After each video, we asked whether something was not clear, providing printouts of the concepts. Then, we probed on participants' first impressions, asking them how they felt when watching the video and what they felt about different aspects of the concept. Following Odom *et al.*'s approach [214], in the final interview, we asked participants to reflect on and across their experiences of all concepts. We asked them to pick the most or least valuable for them, discuss the most positive or negative aspects across all the concepts, and elaborate on the ideas behind them based on how they would fit their needs and experience. Participants used the printouts of the systems to compare and contrast them. One member of the research team conducted all interviews, in English, at our university. Interviews lasted between 37 and 70 minutes (on average: 49 minutes) and were audio-recorded. Participants received $15 as compensation.

7.5.3 Data analysis

To analyze the data, we used Braun and Clarke's approach to thematic analysis [61]. We transcribed participants' answers and started analyzing them inductively using open coding. Then, we grouped codes into categories and developed themes across categories. One member of the research team coded the data and discussed the themes and interpretations with other authors during multiple meetings. We also categorized participants' reactions to each concept as positive, negative, or mixed, and compiled a list of the most and least valuable concepts for each participant.

7.6 Thematic analysis results

In this section, we present the results of our thematic analysis. In general, participants appreciated the idea of getting help in selecting data–they saw it as an important but often challenging task. However, there were striking differences in how they reacted to the concepts (Figure 7.8 provides an overview of participants' individual reactions). For example, reactions to Temporary Folder ranged from enthusiastic (P15: "*I like that one, a lot!*") to perplexed (P16: "*Why would one want a temporary folder?*") and terrified (P1: "*I would be terrified to put something in a folder that's going to be deleted!*"). In the first two themes of the analysis, we contrast diverging opinions on what role technology should have in supporting selection practices: some participants preferred to retain full control of the process (Theme 1). Others, instead, welcomed automation and felt comfortable in offloading selection tasks to technology (Theme 2). Then, we synthesize a middle ground between the two different stances in participants' overall desire for a contextual approach and how this leads to a new perception of keeping and discarding actions (Theme 3). As in Chapter 4, this analysis provides an abstracted, high-level synthesis of cross-cutting individual reactions. The first two themes we present might give off the impression that participants expressed one or the other opinion, but their attitudes and reactions were contextual, as Theme 3 details. In Chapter 6 we already connected these high-level themes to different behavioral styles.

P	Patina	Recommender	Temp Folder	Temp App	Filters	Most valued concept	Least valued concept
P1	mixed	mixed	negative	mixed	mixed	Patina	Temp Folder
P2	positive	mixed	positive	positive	positive	Filters	Temp Folder
P3	positive	mixed	mixed	positive	mixed	Filters - but retrospective	Filters, Temp Folder
P4	positive	negative	mixed	positive	negative	Temp App	Filters, Recommender
P5	mixed	mixed	mixed	positive	positive	Temp App, Filters	Patina
P6	negative	positive	mixed	positive	positive	Filters, Recommender, Temp App	Patina
P7	positive	negative	mixed	positive	positive	Patina	Temp Folder
P8	mixed	negative	positive	negative	mixed	Temp Folder	Temp App
P9	positive	positive	negative	positive	positive	Temp App	Temp Folder
P10	positive	positive	positive	mixed	positive	Filters, Patina	Temp Folder
P11	mixed	positive	negative	positive	mixed	Recommender	Temp Folder
P12	positive	negative	positive	positive	negative	Temp Folder, Temp App	Recommender
P13	negative	positive	negative	negative	positive	Recommender, Filters	Temp App, Patina
P14	negative	negative	negative	negative	positive	Filters	Temp App
P15	mixed	mixed	positive	positive	negative	Temp Folder	Patina, Recommender
P16	mixed	negative	negative	mixed	negative	none	all

Figure 7.8: An overview of participants' varied reactions to the five concepts we created (here abbreviated as Patina, Recommender, Temp Folder, Temp App, Filters). In the first column, the participant number; in the next five columns, participants' general reactions to a concept categorized as positive, negative, or mixed; in the last two columns, the most and least valued concepts for each participant.

7.6.1 Theme 1: Selecting data is a personal responsibility

The first theme captures opposing reactions towards support in selecting data. Some of the participants were generally against automation, mass processing of items, or aggressive systems. Their reactions to the design concepts highlighted a need for control, a sense of responsibility for their data, and a strong desire for doing things in a specific way (the "right way" [288]), all on their own.

Wanting full control

For instance, P11, a professional photographer who managed thousands of photos between her phone, laptop, and an external hard drive, made it clear that fully automatic tools would not work for her because they crossed an important boundary. She felt that having full control over data and the selection process was essential.

Her career depended on properly managing digital data, with no room for mistakes: "*I've heard the horror stories of photographers not backing up data properly and losing up a whole shoot, and, yeah, it will pretty much just ruin your reputation.*" She explained how automatic tools felt intrusive and undermined her sense of control: "*For data, and maybe this is my personality or work, but you don't ever want somebody coming in and telling you 'this is mine'. Or, 'get rid of it'. You make the work, you wanna have control over it. That's why I wouldn't want something like Future Filters going through my files.* [...] *I really I don't appreciate that.*"

Thinking independently

Similarly, P13, who used to work as a "programmer of sorts" in a medical imaging company and often felt overwhelmed with data, highlighted that thinking independently and taking care of items without the help of technology was important to feel in control. In her reasoning, she drew a parallel between some of the concepts and older recommender systems in Word processors (i.e., Clippy): "*It's almost like, you know–there were Word processors that tried to think for you. 'Oh, it looks like you're writing a document, let me do such and such.' And I'm like so mad. I did a lot of Word processing. And I know what I want, I know the spacing I want, I know the editing I want. I know what I'm trying to do and this nuisance tries to think for you. I don't like that.* [...] *Personally, I would like more power and control myself.*" After seeing Temporary App, she added that automatic tools made her feel lazy, hinting once again at the idea that selecting data is a personal responsibility: "*It's kind of like, you feel lazy. Because how hard is it and throw it into the delete stuff? Are we all that busy that we need* [this]?"

Distrusting technology

Underlying many statements from participants there was a sentiment of distrust towards technology. For example, participants questioned how the machine learning from Data Recommender would work and whether it would learn the "wrong things." Similarly, they feared that any function where they did not have full control would eventually go wrong. Thus, a lingering feeling of uneasiness. "*I tend to like the ones that remind you or prompt you vs automatically doing it for you,*" explained P7, an HR specialist who did not trust any cloud platforms for personal data management preferring to do things on "her own," with a mix of collecting and purging. She articulated her preference for Patina and Data Recommender in terms of trust and comfort: "*I wouldn't feel comfortable putting in parameters and just having the technology determine for me. I'd prefer to have them notify me or go through them and choose. I liked* [Data Recommender]. *I'd feel much more comfortable with that vs having things automatically deleted.*"

These preferences were not always a direct reflection of differences in general curation approaches (e.g., tending to keep a lot vs. keeping little). Consider the case of P15, working as an administrator in a government agency: she self-described as a "very organized" person who "doesn't keep a lot of things," limits her technology use, and deletes photos, videos, files, reflecting a purger and frugal approach. She was onboard with discarding data, but trusted herself more than a tool: "*I'm not someone who keeps everything. So I'm not at all reluctant to delete things at once, I know people who are. But at the same time, I don't know if I would trust the computer to delete things if I haven't reviewed them and made sure I want to delete them.*" This explains why she was enthusiastic about Temporary Folder but not Future Filters: "*I have control over what I'm putting in the folder.* [With Future Filters], *you don't really know exactly what it's deleting, you're just trusting that you're putting things in the right place so I feel there's more potential for errors with* [Future Filters]."

Changing idea and having the final say

The need to have full control and think independently informed a strong preference towards always having the final say in all keeping or discarding decisions. Being in control meant seeing recommendations as nothing more than suggestions that need approval and leaving space for changing ideas. When pondering prospective decisions, participants who were generally against automation felt uneasy and wondered what would happen if they changed their mind: "*I think I like the concept of* [Future Filters] *but it's such broad categories that you might end up deleting data and regret it.*" (P8). Anticipating regret, some participants said that keeping everything "just in case" [271] seemed a better approach, while others saw a safeguard in the possibility of controlling decisions and having a final say. For example, P9, a journalist who used to deal with large amounts of data but who wanted to be "more organized" and "have less," explained that "*Across the board, the review process is really important: before something gets deleted I should know it's getting deleted, it should not get deleted without me knowing. And I should have the physical option of choosing to delete or not.* [...] *I should have the final say.* [...] *Sometimes you change your mind.*"

7.6.2 Theme 2: Selecting data is a chore

The second theme captures reactions towards the concepts that contrast those explored in Theme 1. In this case, participants welcomed automation and generally expressed the need for tools that would take care of things for them, freeing them from the weight of selecting data. They felt tired of the responsibility of the selection process, something they put off or were not good at dealing with, and were happy to offload the process to technology.

Being tired of taking care of data

Participants who had positive views of automation were happy about tools taking care of selecting data, a process that they perceived as tiring and relentless. For example, P14, an HR specialist who showed both collector and overwhelmed behaviors, said she "seriously" loved Future Filters and explained that it would help her deal with things she was tired of: "*I am tired of organizing my information and taking care of it every 2-3 months because of the space limit. It bothers me a lot, so if I can set the filters just once for the majority of things that bother me–and it would be pictures, videos, and music, that's the main problem–that would be just perfect.*"

Needing a push to feel the urgency

Participants noted how they needed a "push" to attend to data: "*I thought the most useful* [aspect of the system] *was that it actually popped up. So you have to actually take action on all of the items, because that forces you to decide what to do with them. I thought that was very useful.*" (P10). Selecting items was a task they wanted to engage in, but often put off: "*I feel this could be a cool way to do it for me, because it's something I put off. The last time was probably a year ago or something, so having an app do it for me would be great,*" said P2, who largely had a casual approach to data. More automatic and prospective systems felt useful in creating urgency: "*I think Patina* [was the least valuable]. *Even though it sets an indication, it doesn't create an immediate urgency*" (P6).

Desiring a proactive system

Some participants were satisfied enough with recommendations, but several had preferences for stronger intervention, expressing the desire for a proactive system that would think for them: "*It's perfect if the program can think for me in advance.* [...] *The* [Data] *Recommender is going to bother me for sure. It means the program advises me to think about something and I want the program to think in advance, give me some kind of solution.*" (P14) This preference might have come down to personal style. This is how P10, a student in Education who reported constantly running out of space on his laptop, related tools such as Future Filters to his self-described "lazy" selection style: "*I think it depends how organized you are. Future Filters automatically deletes without telling you what. Patina and Data Recommender remind you and you decide what to do. So, if you feel like you need that reminder and you can delete yourself, I think they would be nice. If you feel you're too lazy or not organized enough, Future Filters takes care of it for you. I am less organized, that's why Future Filters is a really good option for me.*"

Deciding in advance to not worry later

The enthusiasm for simplifying selection extended also to prospective decisions, with participants relating the idea to their own practice: "*There's definitely things I know I don't need. You know, pictures from the internet you want to send someone and they stay on your desktop.*" (P4). Several participants preferred to decide in advance and not worry later: "*I think it's a good idea in terms of coming up with some parameters for things you know you're not going to need in the future and it's better to just automatically delete it and not worry about it*" (P7). They perceived such options as a way to limit the constant input required for selection, a process they compared to daily chores: "[Future Filters] *might be better because I've made a decision and then it will happen and it's not being dependent on me being... you know, it's like cleaning or doing dishes.* [...] *Your input is at the beginning and then it automatically takes care of itself.*" (P13) However, as the next theme shows, these decisions still needed to have some safeguards in place or otherwise respect the context of data.

7.6.3 Theme 3: Context is key

In Themes 1 and 2, we have described a range of reactions to the concepts, with two general contrasting stances. The final theme highlights how these reactions were different but never completely polarized, because the context of data played a key role. Participants noted important differences between data types based on their nature (a document you take a lot of time to create vs. a movie or an app, that you can always download again), their context (work data being generally more important and critical than personal data), and the device under consideration (computers being for serious stuff and smartphones being for less critical stuff). They perceived data as being always somewhere out there, in the cloud or on a device: this had both positive and negative consequences, and informed what we call a *post-cloud* perception of selection.

Selection decisions are contextual

A recurring thread in participants' reactions, whether more positive or negative, was that keeping or discarding decisions are highly contextual. Thus, a concept that worked for one type of data, might not have worked for others. For example, many participants drew distinctions between work and personal "stuff," saying that they tended to be more organized and less selective with what to keep at work: "*I keep everything for work related, for personal it's different.*" (P7) Similarly, they regarded data on phones as easier to discard and often less important. They perceived smartphones and mobile devices as "fluid, temporary, and more accessible," compared to laptops, that were "serious, demanding," and with more places where to hide things away. The difference in storage capabilities between the two types of devices was also an important factor to consider: "*For the phone I periodically* *delete pictures I don't want and apps I don't use anymore. But on the phone it's partly due to storage problems. Which is not that much of a concern with my Mac,*" recollected P3 in illustrating their largely purger approach.

Exceptions to general decisions

Participants also remarked that digital data being old or unused did not necessarily mean that they would have liked to get rid of it, as some concepts suggested: "*I don't like the fact it says you haven't used it because it might say, you haven't used it in six months, get rid of it. But that's not a good idea because sometimes you save files for you future situations. Deleting files because they are older, is that a good idea? Maybe there's a reason they should be kept.*" (P16) They wanted to define exceptions to general decisions and have the option to instruct the system about any item that they might want to keep: "*Maybe there's an option to exclude certain things. Like, all photos that are older than 2 weeks, except these three. That'd be a good option to have, to be able to create exceptions.*" (P10). These reactions point to the importance of marking items to keep explicitly, an action often absent from data management tools.

The cloud is as big as the universe

The contrast between different contexts and storing places was particularly evident when comparing Temporary Folder and Temporary App. Several participants explained that automatic or prospective actions were more acceptable with mobile applications because applications are not unique and are not the result of time or effort: "*There's no big risk, you can install the app again*" (P5). When expanding the focus within their data ecosystem [288] and discussing the cloud, with places such as Facebook, Google Drive, or Dropbox, participants noted how the change in context changed their attitude, hinting that selection would be less necessary in these storing places: "*I never delete* [from Facebook] *because I imagine their storage is as big as the universe.*" (P6)

Data is always somewhere out there

The key role of cloud platforms and the interconnected nature of data ecosystems informed a perception of data as ubiquitous and perennial. Participants in Gulotta *et al.*'s study [111] saw deleting as being against the nature of digital data. Participants in our study perceived data as never truly deleted because there will always be a copy somewhere, out there. Perhaps it is a copy on an external hard drive, maybe it is a backup on Facebook or Google, but data never really disappears unless you want it to: "*We've talked about deleting apps on my phone regularly and things and as long as they're backed up somewhere, if they're deleted it's not a big deal.*" (P2). This *post-cloud* conception of data makes keeping and discarding decisions take on a different meaning: deleting means removing data only from one specific device or removing a specific instance, while having a copy somewhere else: "*I don't think I will put anything in Temporary Folder that I don't have a backup for, so it's fine.*" (P8). And archiving really means moving or hiding data within a device ecosystem: "*I wonder how the archive... maybe it's like moving, it sounds like a similar function. I* [archive] *with my emails, but I think it goes to... it's the same idea as moving.*" (P6). Suddenly, automatic or prospective decisions are more acceptable: "*Especially in the selfie scenario, you probably already posted it on Instagram or Snapchat or whatever, so there's a copy of it already in the world, so removing it from your device it's not a big deal.*" (P5). The selection process then becomes a matter of moving data back and forth within an ecosystem and the cloud is the ultimate storage utility.

7.7 Discussion and future directions

7.7.1 Moving the design space towards personalization

The range of reactions we gained from participants supports the idea that decisions around what data to keep or discard are highly personal, as we have seen in Chapter 4 and Chapter 6. As we expected, no single solution was able to resonate with most participants. But these results show how branching into potentially controversial or radical areas of the design space can be fruitful. By inquiring into concepts that appeared risky, we were able to get a better idea of where people's boundaries lie, what it means to cross them, and how different people may have different boundaries. The design dimensions we explored can now be used as a generative resource to work towards new solutions. Some of the design dimensions and concepts we explored could be remixed (e.g., providing a list of filters that can feed into recommendations) and modulated to support different user attitudes. There is an opportunity for future work to further investigate this emerging space through designing, developing, and studying more personalized solutions (e.g., customizing default keeping and discarding actions or criteria for recommendations). In the following sections, we articulate some more key directions for future efforts.

7.7.2 Finding a space for automation

In our analysis, we were particularly fascinated by the contrasting attitudes towards automation. Do the strong reactions against some of the concepts mean that we failed as designers to support user needs? We think the key here lies in unpacking the underlying threads of such negative opinions and leverage them to move towards a more nuanced approach. As highlighted in the related work, the tension between automation and user control is a long-standing issue. A key contribution of our work is revealing that some keeping decisions, under specific circumstances, can likely be automated. In particular, promising initial contexts to pursue automation in design interventions are mobile devices with limited storage space, media files, and distributed data (i.e., data that is not unique or otherwise re-accessible). In other cases (e.g., different devices and types of data), there still is a space for automation, but only with proper safeguards. This principle extends to other dimensions of the design space, as we outline next.

7.7.3 Safeguarding automatic and prospective decisions

Our analysis reveals that there can be a space for both *retrospective* and *prospective* actions, *manual* and *automatic*, and *open-ended* or *specific*. However, it is essential that future design interventions synthesize these extremes so that any action is reversible and any potential risks are mitigated in advance. This suggests an opportunity to investigate how to design effective safeguards for automatic and prospective actions. The easiest way to design safeguards would be to provide reminders before automatic or prospective actions, something that several participants asked for. Another approach would be to simply promote softer actions over the more radical concept of deletion (e.g., moving, trashing, hiding). Yet another opportunity would be to see the perceived risk and anticipated regret that come up in participants' words as explicit components of the decision process. For example, systems could visualize a history of prospective or reverted decisions (e.g., how many times a document was marked for trashing and then reverted, or how many times a mobile app was uninstalled and then re-installed over the course of a period of time). Similarly, efforts along the other dimensions of the space could focus on letting users explicitly define potential risks and regrets and then evaluate them at a later point in time.

7.7.4 Rethinking keeping and discarding actions

When reacting to the concepts and the actions they afforded, some participants struggled with understanding what an archive is. For others, archiving was the same as moving or hiding. Similarly, participants reported deleting practices rooted in the importance of context and the availability of a multitude of platforms and devices. Sas *et al.* [249] argue that "deletion is a crude binary process," while Ramokapane *et al.* [240] highlight how cloud platforms in particular provide poor deletion models. We agree that a binary representation of data as either present or deleted does not reflect the majority of our participants' mindset. This argument ties to *prospect theory* [145] and can further explain why keeping decisions are so challenging: if deletion is a binary process, it is more difficult to balance risks and gains. Yet, in most tools deleting is the default discarding action. Although our results show that crude deletion is welcomed in some cases (e.g., with mobile applications), moving towards a mitigated process of deletion might be the way forward to support different contexts. This idea resonates with work by Harper *et al.* [123] who argue for rethinking actions about owning, copying, and deleting data. Based on these implications, we see two possible directions to follow. The first is a design-focused effort in the line of work by Lindley *et al.* [167] to explore and define a new grammar of actions around keeping decisions. Harper *et al.* [123] propose "eradicating" or "withdrawing" files from the cloud, while Bergman *et* *al.* [29–31] show examples of "demoting." This set of actions could be extended to include *mirroring* (for storing a copy of data from a central repository only temporarily), *distributing* (to disseminate copies of data around many storing places), *warranting* (to authorize automatic tools to act only on specific items), *locking* (to mark items as protected by any discarding action). Providing options for discarding actions to be granular rather than binary will also allow for more personalized solutions. For example, some users might choose "crude" deletion, while others might opt for demoting or distributing items as their primary selection action. A second, research-focused effort would be to further disambiguate actions as used in interfaces and perceived by users with a taxonomy. This would follow the example of Watkins *et al.* [299], who categorize very specific different types of digital "collections" and collecting practices. Research often refers to the old metaphor of an archive to study and discuss people's practices. It seems like apt timing to take more seriously the question of what is an archive in the post-cloud age and is this the best metaphor to use?

7.7.5 Taking steps towards active data privacy protection

A tangential but important issue that came up in our exploration was the topic of privacy and security. This was not our focus, but we inevitably touched on it. Participants discussed how the concepts could work for managing privacy and security, both on their devices and in the cloud. Their attitudes varied. In general, they perceived the concepts as acceptable if they came from trusted brands, were officially part of the operating system, or were looking at data on devices more than on cloud platforms (possibly because cloud platforms are "curated through use" [315]). But often participants noted how keeping decisions are more delicate and consequential when it comes to privacy. Nudges, reminders, and prospective actions could prevent unwanted issues or "leaks" of sensitive data.

These results resonate with the public's desire for more control over data [176] in the face of recent data scandals [47, 107]. The advent of the cloud has imposed a centralized data management model where a few corporations (Amazon, Apple, Google, Facebook, Microsoft) aggregate the large bulk of people's data. But as Mortier *et al.* argue [198], this approach is "fundamentally flawed" and as the importance of digital data continues to grow, it faces increased scrutiny. We argue that a contextual and user-driven approach to keeping decisions in the cloud can be a concrete step in protecting privacy. In particular, our ideas around safeguarding and rethinking keeping decisions can be extended to a privacy-oriented mindset to provide more control to users. There is an opportunity for future work to target similar ideas exclusively around data privacy management and better explore people's attitudes. This approach would fall in line with recent work on "design workbooks" for privacy [312]. Another possibility in this domain is to study how to use similar concepts for data created *about* people (e.g., advertising data). Opportunities include allowing users to create temporary advertising profiles or review and discard any tracking data that companies have on them.

7.7.6 Reflecting on the broader impact of our work

Finally, we reflect on the broader impact of our work for individuals and society [125]. On one side, we hope to inspire positive change in the space of data management and selection, pushing towards a nuanced approach and truly user-centred designs. At the same time, we see how some of the concepts we propose might lead to unintended consequences and be abused to further centralize data management and restrict users' freedom. For example, Temporary Folder and Temporary App could be used to restrict users' access to their own cloud-stored data, imposing a subscription model to items that they perceive as their own. Indeed, we see this trend already emerging in several software applications, as P11 lamented when discussing her use of the Adobe Creative suite. Business needs drive these decisions, yet, this approach may contribute to eroding people's sense of ownership and agency in relation to their data. This, in turn, feeds their general distrust for technology. We argue for an alternative, de-centralized data management model, where users' control is key, privacy comes first, and management actions are context-based. New regulations such as the European Union's GDPR (General Data Protection Regulation) provide a first step towards more ethical practices in the space of personal data, but regulatory efforts need to be complemented by design changes. It is up to us, as researchers and designers, to ensure that the needs we discuss around personal data are met and that people's boundaries are respected.

7.8 Limitations

Some limitations in our study point to additional future work. First, while our sample is meant to be generative and varied in terms of occupations and data management styles, participants had a predominantly Western background. This limitation is an opportunity for future work to focus on participants from different cultures, to see if and how attitudes change. We also did not screen participants for more general attitudes around decision making or psychological traits, because finding correlational links was not the goal of this study. But there might be individual differences in risk aversion and risk seeking that inform people's attitudes in deciding what data to keep or discard. A previous study by Massey *et al.* [187] links personality traits to differences in file management behaviors. Similar efforts along this line can complement our work.

7.9 Summary and conclusion

Drawing on previous work on personal data management, we created five concepts to explore a design space around keeping and discarding decisions. By probing different design dimensions, we elicited contrasting attitudes about the role of technology in supporting decisions, finding a common ground in the need for nuanced and contextual support. Our work in this chapter opens possibilities for new tools that, with proper safeguards, have the potential to help users better select what personal data to keep or discard. We see this as a critical step in addressing our *post-cloud* world that is overflowing with data.

In the next chapter, we will explore how to include ideas and implications from this study into a cohesive, centralized system, with a focus on personalization.

Chapter 8

Study 4

Evaluating a Personalized Approach to Personal Data Curation

In this final research chapter[1], we build upon the results of all previous chapters. We use the insights on preservation tendencies (Chapter 4), the five curation behavioral styles (Chapter 6), the taxonomy of decluttering criteria (Chapter 5), and the insights on selecting data (Chapter 7) to build *Data Dashboard*, a cohesive, personalized system that combines different design ideas from across the dissertation. We then use a prototype of the system to explore how this approach can support people's curation behaviors. Our results are promising and provide support for much of the work that came before, but they also leave space open for future work in this area.

8.1 Introduction

Throughout the dissertation, we have shown how deciding what personal data to keep or discard can be difficult. Selecting what to keep is a contextual choice and data management tools often offer poor support for this practice, making personal

[1] Published in Vitale, Chen, Odom, and McGrenere. (2020) *Data Dashboard: Exploring Centralization and Customization in Personal Data Curation*. Proceedings of the 2020 Conference on Designing Interactive Systems (DIS '20) [292]

data curation a challenging process. In this chapter, we focus on what we think are the two most pressing challenges that complicate user practices and possible design efforts that might mitigate them.

The first challenge for personal data curation is the growing number of devices and cloud platforms that people use [288]. The distributed, often fragmented nature of personal data undermines awareness of ownership, making curation even more difficult: it becomes hard to curate data if you do not know what you have [213]. One approach to address this issue is ***centralization***, which we define as providing an overview of data from different places in a single central tool for increasing awareness of what a user owns [213, 215]. Our work asks the following questions: Can centralization help people decide what personal data to keep or discard? If so, how should we approach the design of centralized tools? With cloud platforms such as Google Drive and Dropbox moving towards centralization by encouraging users to "sync" all their data to the cloud as a backup, investigating this question can produce insights that will help us understand the consequences of a similar approach.

The second challenge for building curation tools is the subjective nature of personal data management and curation [22, 219]. It is difficult to create a solution that can satisfy different types of users who have different management styles and curation approaches [150]. Our work in previous chapters details the subjective nature of keeping and discarding decisions. A common design approach to deal with individual differences is to turn to ***personalization* or *customization***–the ability to tailor a system to specific users' needs and characteristics. From this comes our second set of research questions: Given the personal and subjective nature of data curation, can customization help? How desirable is a personalized approach to data curation? And, what are the aspects that make it more or less desirable?

To answer these questions, we designed Data Dashboard, a centralized and customizable system for curating personal data. We evaluated an interactive prototype of Data Dashboard with 18 participants who had different approaches to data curation, asking them to go through five potential use scenarios. Drawing on previous work around digital data, we use the concept of ***data boundaries*** (the idea of conceptual lines that prescribe where to store personal data and how) to understand participants' reactions to centralization and customization. We show that centralization *blurs* boundaries and introduces a dilemma around privacy and security, requiring explicit safety guarantees. Customization, on the other hand, is easier to accept because it *upholds* boundaries. We discuss what these results mean for future data curation tools.

In this chapter, we make three contributions: 1) we provide additional empirical evidence for the role of data boundaries in personal data curation and use it to understand participants' reactions to design choices–specifically, we show reactions to centralization and customization; 2) we offer an approach to address key challenges in data curation by designing a unified tool with personalized functions; and 3) we outline design and research directions for future data curation tools focused on integrating data boundaries into design, rethinking the language of personal data, and envisioning a post-cloud future.

8.2 Related work

The related work we discuss in this section falls into three main areas: (i) previous PIM system studies; (ii) data boundaries; (iii) personalization and customization. Additional related work appears in the prototype description.

8.2.1 Previous PIM system studies

Several PIM studies propose and evaluate new tools for managing data. For example, some systems focus on management and retrieval of documents [69, 79, 82, 90] or files and folders on desktop computers [29, 143]. Others explore how to manage and curate photos [32, 208, 245, 320], contacts [30], or emails [19]. A related strand of work explores automation in the context of task management [119] or cloud file systems [34], showing that it can help people in their management tasks. However, attitudes towards automation for data curation show individual variation, with some people opposing automation (as we show in Chapter 7).

The majority of past design work in PIM takes a quantitative approach for evaluating prototypes, using lab experiments or usage log analysis in field deployments. These studies largely look at time on task and similar metrics for measuring success. A few studies, instead, use a more qualitative approach, focused on teasing out participants' attitudes. In our work, we use a qualitative approach in line with

Research through Design [317] and reframe design artifacts in Chapter 7 and Chapter 8 as tools to understand broader aspects of data curation. This approach falls in line with past work on digital data that uses systems to prompt discussion with participants [111, 112, 212, 232]. Below, I expand on my methodological approach throughout the dissertation.

8.2.2 Data boundaries

Past work points to the idea of *data boundaries* as a key lens to understand data curation. At a high level, we define data boundaries as conceptual lines that prescribe where to store personal data and how. Boundaries appear in studies on data management [219, 288, 301], collaboration [293, 295], communication [49, 51, 207], privacy [2, 15, 225, 231], and personal possessions (both physical [17, 58, 77, 270] and digital [154, 166, 211, 299]). We can argue that data management and curation are essentially about establishing and negotiating boundaries. This process has roots in cognitive models of how the human mind works: to make sense of a continuous world, people build categories and "draw mental boundaries around them" [219].

Research on digital data often looks at the contrast between physical and digital possessions, exploring the boundary between the two domains [108, 111, 140, 149, 153, 211, 215, 224]. The notion of boundaries runs throughout studies on physical possessions, helping us understand how people might experience them. For example, people might use boundaries to mark the unique character of their spaces [77], give meaning to clutter [270], and establish what belongs in their home [58]. These are all practices that help in building an identity [17]. Boundaries also exist in the digital world. Sometimes they are explicit, as is the case when people use folders to create structure in their data [301] or when they build collections with clear criteria for what goes in and what does not [299]. But more often data boundaries are implicit. They are influenced by tools and applications [19, 166, 288, 293], work and personal life [49, 51], group and family relationships [154, 207, 288, 296], activities [295], and context (Chapter 7). A key boundary to understand participants' reactions to centralization is the one around privacy. Previous studies argue that privacy management is at its essence about negotiating personal boundaries between the public sphere and the private sphere [2, 225, 231]. Privacy boundaries are "permeable" and "murky" with context playing a key role in defining and shaping them [2]. They are subjective and dynamic, evolving over time [15, 225, 231].

Our work provides additional evidence for how boundaries drive personal data curation and shows how to use the concept to understand reactions to centralization and customization. We did not have in mind the concept of boundaries when we designed Data Dashboard or the study, but we identified it in the analysis and then traced it back to past work.

8.2.3 Personalization and customization

Broadly speaking, personalization is about creating a system that is adapted to the user's individual preferences and characteristics [12, 36, 119, 190, 193, 269]. There is no standard definition of personalization, but previous work highlights different levels of user involvement [36, 190, 193, 269]. In purely "adaptive" system-controlled personalization, the system does not directly involve users when deciding how to change the content, interface, or functionality [269]. System-controlled personalization is largely implicit. Customization, by contrast, is the term used for user-controlled personalization, also sometimes referred to as "adaptable": the user is responsible for changing the system through explicit actions. A middle ground approach involves mixed-initiative systems, where personalization is system-initiated but needs to be "approved" by the user [46, 121, 134].

Regardless of level of user involvement, personalized changes in the system can take place at different levels: user interface (layout), content, and functionality. Changes in the user interface include, for example, changing colors, fonts, backgrounds, visible buttons, and so on. These changes are largely concerned with aesthetics [314]. Changes in content are relevant in information-centric systems: they involve showing or hiding specific content based on different user needs. For example, social network feeds show different content to different users [190, 269]. Finally, changes in functionality are about changing how the system works [119, 314]. For example, creating macros in spreadsheets, extending browser functionality with extensions, or changing system settings. This approach is often labeled “advanced customization or personalization” [119]. In this study we focus on content and functionality changes.

Personalization can help users complete tasks and provide benefits if done well [46, 92, 106], but there are factors that influence and limit its benefits. In general, users seem to prefer mixed-initiative interfaces over purely adaptable interfaces [46], and adaptable interfaces over purely adaptive interfaces [193]. But the time and knowledge required for enacting customization often prevents people from enjoying its benefits [174]. Individual differences [193] and factors such as exposure, awareness, and social influence also determine whether users will customize or not [12, 92]. Several PIM studies have looked at automating information organization, with past work [121] (and Chapter 7) providing a more detailed overview of related work in the area. However, only a few studies have looked explicitly at personalization or customization mechanisms for data management [119] and curation [208, 245], suggesting that this is a promising but unexplored approach.

8.3 The Data Dashboard prototype

We designed Data Dashboard to address the two key challenges of personal data curation outlined in the introduction: the fragmented nature of data across devices and the subjective nature of curation.

8.3.1 Overview of the prototype

Data Dashboard is a centralized system that provides an aggregated overview of data stored on different devices and cloud platforms (e.g., Dropbox, Google Drive, iCloud). The system provides an overview of data and a set of customizable filters for curating data. There are four sections: Activity, Explore Your Data, Quick Actions, Settings.[2]

[2]The prototype is online at https://datadashboard.github.io. It works on all desktop browsers, but has some bugs in Safari. Chrome is the best browser to use it. The prototype is not optimized for mobile devices. We note that the prototype is only an illustration of possible functions: it does not connect to any device or cloud platform. A video walkthrough of the prototype is available in the supplementary materials to the thesis.

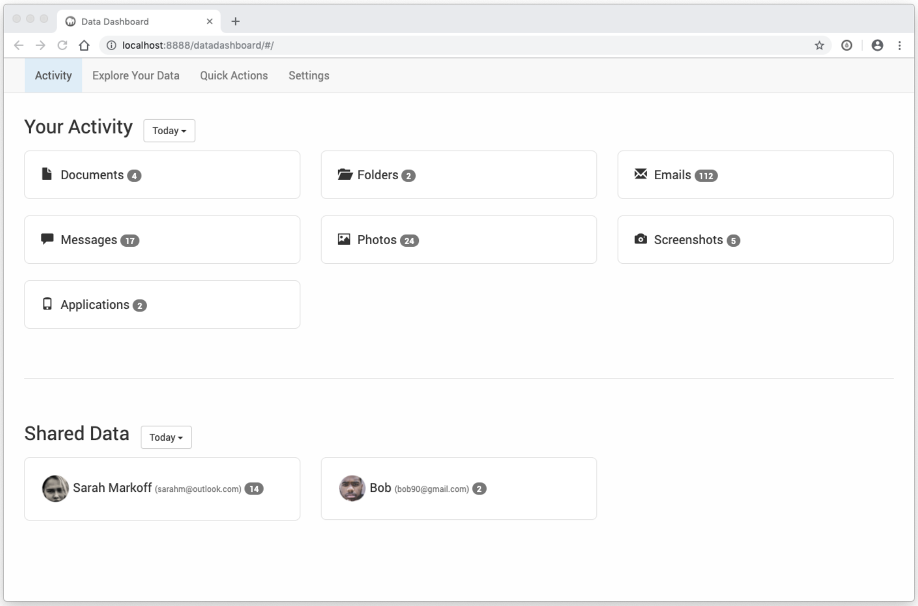


Figure 8.1: Activity shows an overview of recent data.

Activity (Figure 8.1) provides an overview of recent data (Google Drive, Dropbox, and macOS provide similar views of recent files). For example, for a given day, it shows documents that users have created or edited, photos and screenshots they have taken, apps they have installed, and so on. Users can filter their activity by time (today, this week, this month, all time, custom period). Activity also has a section aggregating data shared with other people on collaborative services (e.g., Dropbox, Slack), grouped by the person it was shared with.

Explore Your Data (Figure 8.2) shows an overview of all data users have on different devices and platforms, grouped by type of data (e.g., photos, emails, text documents, videos, and so on)[3]. Users can sort or hide the different data types

[3]The complete list of data types we considered is visible in Explore Your Data after clicking the "Sort or hide the types of data you see" button. It includes: photos, screenshots, emails, text documents, presentations, spreadsheets, videos, audio, folders, contacts, applications, messages, cache and logs, bookmarks, ebooks, tracking data, browser history, passwords, notes and reminders, games. This is a comprehensive but not exhaustive list of types, meant to be a starting point in our exploration. Future work could extend our approach to a wider range of types (e.g., unintentionally digital traces collected by technology, such as the time spent on a device, clicks, queries.)

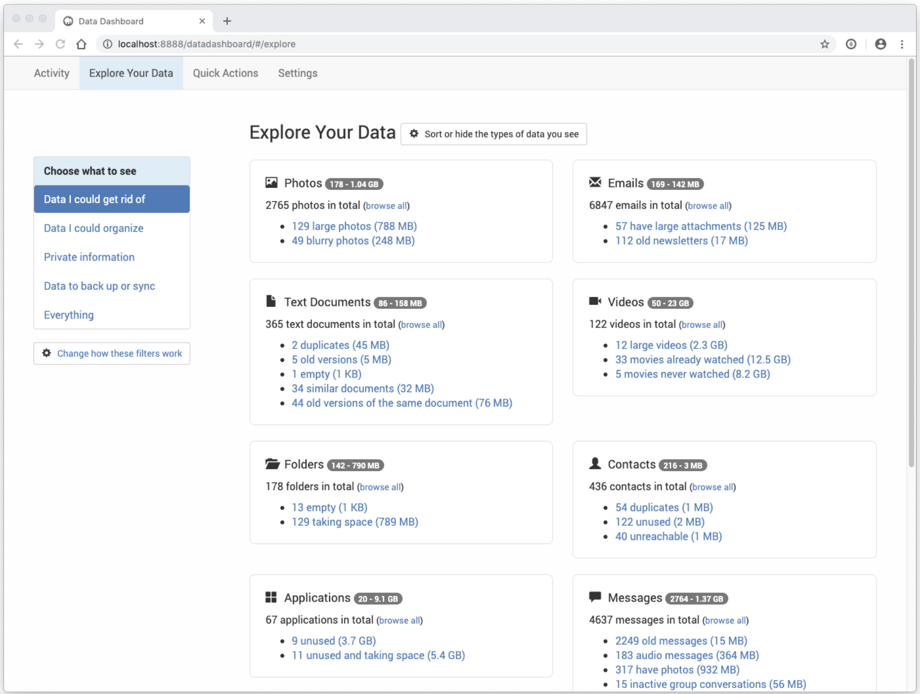


Figure 8.2: Explore Your Data shows an overview of all data from different devices and cloud platforms. Users can use filters to see different categories of data for each data type.

displayed. They can also filter what the system shows them, choosing a filter on the left: "Data I could get rid of," "Data I could organize," "Private information," "Data to back up or sync." Users can customize how the sidebar filters work in the Settings section.

Quick Actions (Figure 8.3) provides a list of recommended actions for different data items. For example, the system suggests duplicates to remove, documents to rename, or message conversations to archive. As in Explore Your Data, users can filter the recommendations using the sidebar filters (same options as in Explore Your Data). Users can also apply automatically suggested actions to similar items in the future.

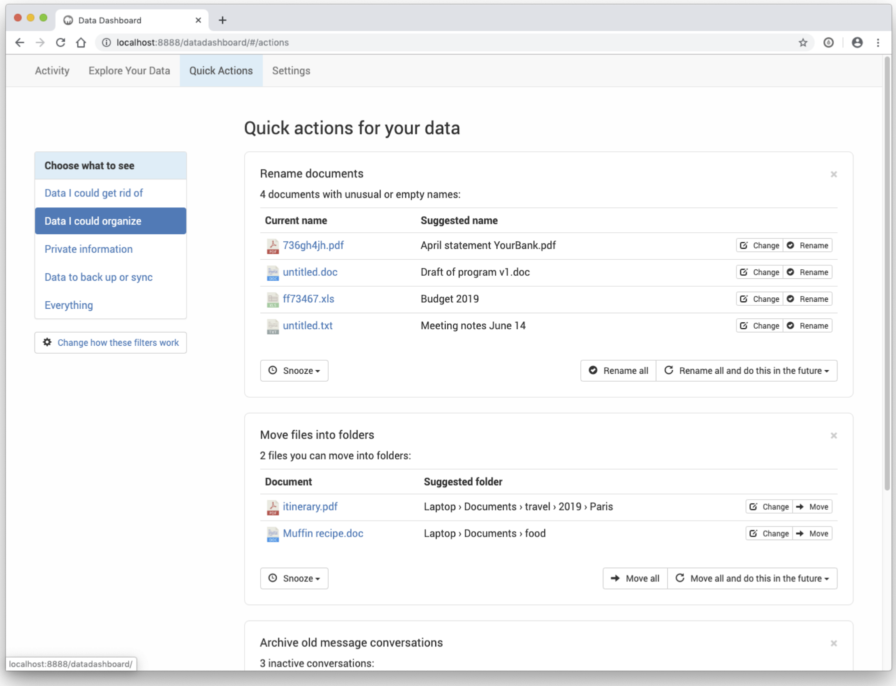


Figure 8.3: Quick Actions has recommendations for curating.

In **Settings** (Figure 8.4), users can customize how the filters for Explore Your Data and Quick Actions work. They can choose to include or exclude some default combinations of data types and criteria or create new, custom combinations with different data types and criteria. They can also add or remove connected devices and cloud accounts.

8.3.2 Rationale and design

Data Dashboard combines insights from the previous chapters and past work on personal data management [166, 213], while also combining and extending key aspects of existing commercial and research products. The visual design of the system takes direct inspiration from Google Dashboard [4] and similar privacy dash-

[4]https://myaccount.google.com/dashboard

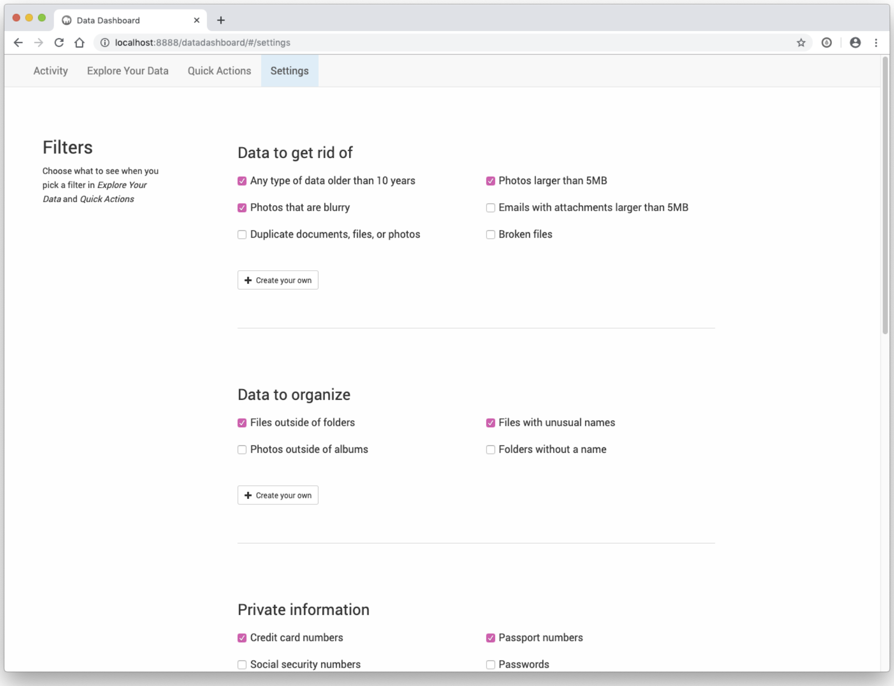


Figure 8.4: The Settings page allows user to customize data filters.

boards. We wanted Data Dashboard to feel similar to existing systems so that participants could better imagine when to use it. Still, we had to build our own system rather than use existing tools to explore how centralization and customization could work together.

The idea of a centralized tool

One key inspiration for Data Dashboard is work by Odom *et al.* [213, 215] on cloud platforms and the need for more awareness of digital possessions. Odom *et al.* suggest creating a "visual inventory" of digital possessions, "a place where 'my stuff' can be found, even if, in technical terms, it exists on many different servers, or many applications." A place to quickly go back where data is originally stored, preserving its context. Lindley *et al.* [166] explored a similar premise and found that a centralized web archive might not be the ideal solution for users. Data Dashboard, however, is not a central archive in itself, as it only provides links to data stored on different devices and cloud platforms but does not copy any data from other storing places (technically, Data Dashboard would rely on metadata provided by devices and cloud platforms to provide an aggregated overview of different items). Instead, we see it as a tool to curate a "meaningful archive" [183] within the systems that people already use to store and manage data.

Some past PIM projects also explore the idea of centralizing data [69, 82, 143]. But they focus on retrieving data, rather than on helping users curate and decide what to keep or discard. Cloud platforms such as Google Drive and Dropbox, instead, are moving towards centralization by encouraging users to "sync" all the data from their computers to the cloud as a backup [81, 246]. There are also commercial products that offer to unify separate cloud accounts, such as odrive [5] and MultCloud [6], but Data Dashboard has specific novel functions for curation (e.g., the data filters and their customization).

Providing a dashboard

A key function of Data Dashboard is to provide an overview of personal data by showing numbers for different categories (e.g., 49 blurry photos). Two systems inspired this approach: 1) Cardinal [75], a research tool that scans a user's computer and provides counts for the total number of files, breaking down numbers for some popular file types (e.g., photos). We wanted to use the same approach in a user-oriented system. 2) Google Dashboard [7], a privacy-oriented page that provides an overview of data created and stored in different Google products [233]. (Yahoo [8] also provides a similar interface for managing privacy settings, and work on GDPR compliance also proposes the idea of a dashboard [241].) Data Dashboard is similar to these tools (Figure 8.5) but extends to entire personal data "ecosystems" [288], bringing together data from more than a single device or platform.

[5] https://www.odrive.com
[6] https://www.multcloud.com
[7] https://myaccount.google.com/dashboard
[8] https://yahoo.mydashboard.oath.com

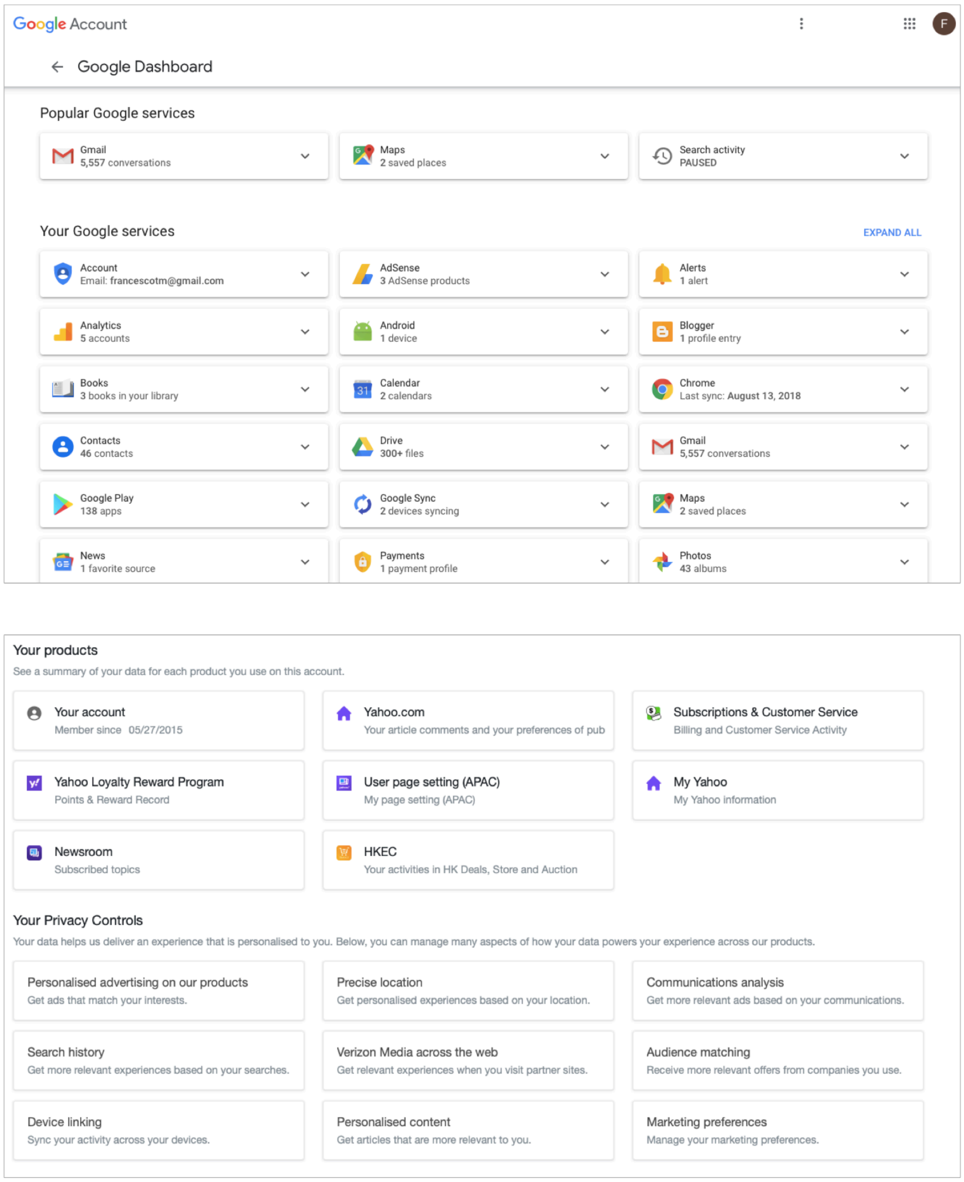


Figure 8.5: The design of Data Dashboard took inspiration from similar dashboards available online. Here, two key examples: Google Dashboard (top) and Yahoo's privacy dashboard (bottom).

Our work in Chapter 7 and "cleaning" tools such as Files [9], Clean My Mac [10], and CCleaner [11] inspired the automatic recommendations for data to curate in Quick Actions. However, we wanted Data Dashboard to be more comprehensive than similar tools. Most of these tools focus on freeing up space by finding items to discard based on their size. Instead, we wanted to help users go beyond freeing up space, and provide them with a broader set of actions.

8.3.3 Personalization mechanisms

Data Dashboard provides both content customization and functional customization. In terms of content, users can customize what data types they see and their order in Explore Your Data (Figure 8.6). In terms of functions, users can personalize how the sidebar filters in Explore Your Data and Quick Actions work, deciding what the system will show (Figure 8.4). This approach is in line with work on advanced personalization for task management [119]. The four options in the sidebar filters ("Data I could get rid of," "Data I could organize," "Private information," "Data to back up or sync") reflect different behavioral patterns and goals that people might refer to when curating data, based on the insights we presented in previous chapters. In particular, "Data I could get rid of" embodies user goals from three key behavioral styles presented in Chapter 6: Overwhelmed, Purger, Frugal. "Data I could organize" is largely targeted towards the Overwhelmed behavioral style, but is also consistent with the cross-cutting need for re-organizing data in other behavioral styles (Casual, Purger, Frugal, Collector). "Private information" is targeted at the cross-behavioral style goal of protecting data privacy, with some styles (Collector, Purger) reflecting a stronger concern compared to others (Casual, Frugal). Finally, "Data to back up or sync" is largely targeted at the Collector behavioral style, with this pattern reflecting a common worry of losing data.

In addition, the three main sections (Activity, Explore Your Data, Quick Actions) target different types of users and data curation practices. Activity and Explore Your Data target users who prefer to explore data and decide on their own what to do by inspecting items individually. By contrast, Quick Actions targets

[9] https://files.google.com
[10] https://cleanmymac.com
[11] https://www.ccleaner.com

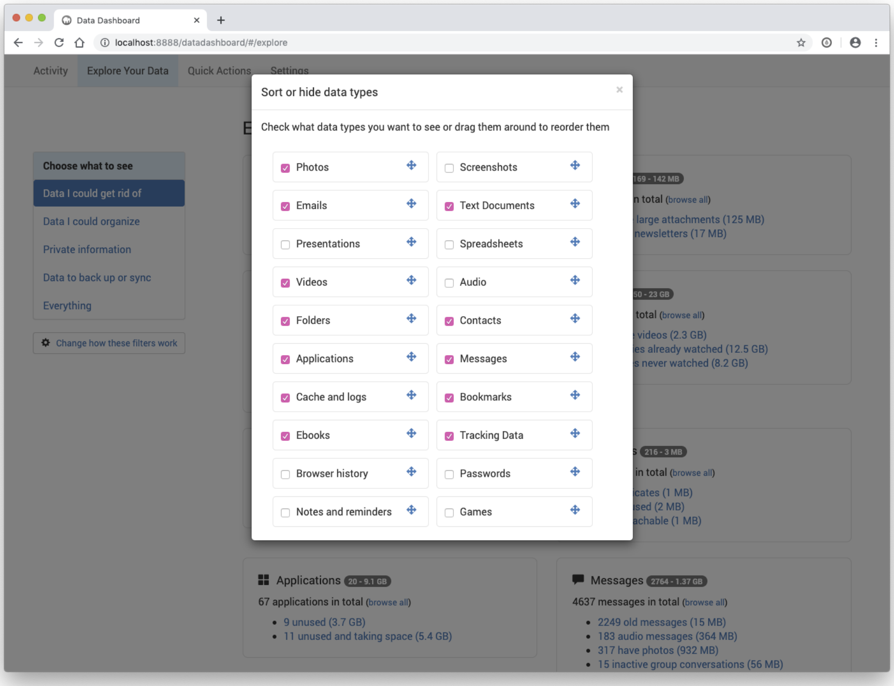


Figure 8.6: The panel to sort or hide data types in Explore Your Data.

users who welcome automation and expect an intelligent system to do things for them. Chapter 7 details these contrasting attitudes to data curation.

8.3.4 Implementation

We built the prototype using AngularJS and Bootstrap 3. We designed Data Dashboard as a system that would be quick to turn into a horizontal prototype, that is, a prototype where only top-level functions are implemented to communicate the scope of the whole system [21]. All data in the prototype is "fake." This is a limitation of our approach, as previous work shows the value of using real participants' data [111, 112]. However, we took this approach because it was easier and faster to prototype. Not using participants' real data also prevented any potential infringement of participants' privacy and gave all participants the same experience when going through the scenarios. The recommendations in Quick Actions, the data types in Explore Your Data, and number of items for each type are meant to illustrate the scope of the prototype and largely reflect common distributions of personal collections [73, 76].

8.4 Methodology

We evaluated Data Dashboard in a user study with 18 participants. In the evaluation, we collected participants' opinions about the Data Dashboard interface, the key ideas behind it, and possible scenarios of use. This specific approach to Research through Design aims to use design artifacts and devices to frame and open prospective conversations with participants, as previous studies show [111, 212, 232, 317]. This approach also explains the relatively minimal, under-designed nature of the prototype: we did not design a costly, high-fidelity prototype because our goal was a design exploration focused on eliciting participants' reactions. Before recruiting participants we ran four pilot sessions to check for potential issues in our study procedures. We also gathered feedback from other lab members throughout the iterative development of the prototype, going from paper sketches to the interactive version we implemented.

8.4.1 Participants

We recruited participants using a university recruiting list and Craigslist in Vancouver, Canada. We used a screening survey where we asked participants their age, occupation, main approach to data curation (based on Chapter 6), and what devices, cloud platforms, and specific data management tools they used (all from a list of popular options). (The screening survey is available in Appendix D.)

We received 169 responses to the screening survey. We contacted 38 respondents, 25 agreed to take part in the study, and we ran the study with 18 of them (12 women, 6 men, aged 18-64, median age: 33). Participants' occupations included accountant, background actor, business contractor, childcare provider, facilitator, occupational therapist, sales associate, social worker, student, postdoc, research manager, retired. Most participants self-reported having average technical skills and no experience in computer science or programming. We recruited for a varied set of participants who used different data curation tools, cloud platforms, and devices. Some participants did not use any tools or cloud platforms. Participants also varied in their approach to curation, based on the behavioral styles we introduced in Chapter 6. We had a roughly equal distribution of participants across four behavioral styles: overwhelmed, collector, purger, casual. Only one participant displayed a frugal approach for some of their data, but we have already discussed in Chapter 6 how we expected a much smaller number of participants with a frugal approach.

8.4.2 Procedure and data collection

The data collection took place over one month, with each study session lasting between 41 and 97 minutes (average: 64 minutes). Whenever possible, two members of the research team conducted the session. One member would ask questions, while the other would take verbatim notes on a computer. In cases where only one member of the team conducted the interview, we later transcribed the audio recording. After each interview, we compiled a debrief document noting key answers and preliminary insights from the interview.

The session had three parts (each lasting 10-30 minutes): an introductory interview, an exploration and scenario-based interaction with the prototype, a debriefing interview with two, short card-sorting activities. Participants interacted with the prototype using the Chrome browser on a MacBook Air 13" laptop. We audio recorded the whole session and screen-recorded the interaction with the prototype. Participants received $20 in compensation at the end of the session.

Introductory interview

In the first part of the session, we had an introductory interview focused on data curation practices. First we asked participants to remember and tell us about the last time they decluttered some of their data, by reviewing, organizing, discarding, archiving, or moving several items at once. Then we asked participants to show us examples of how they organized their data on their devices, how they decided what to keep or discard, and how they used specific tools –if any– to manage or curate their data (e.g., settings panels to clean up data, Google Dashboard, and so on, depending on what they mentioned using in the screening survey and in their interview answers).

Use scenarios

In the second part of the session, we introduced participants to the prototype and let them familiarize with it for a few minutes, prompting them to think aloud. After this initial exploration, we asked participants to go through five possible use scenarios with the prototype, once again asking them to think out loud.

Scenario 1 (space running out): “The space on your computer is running out. You want to find some data to discard. You are not sure where to start looking, but you know that you do not care too much about old documents.”

Scenario 2 (taking time for regular data curation): “It is a rainy day. You have set aside some time for doing a regular cleanup of your devices. You usually do this every few months. You want to review your data and make sure everything is organized in your preferred way.”

Scenario 3 (exploring recent data): “You have 5 to 10 minutes in between meetings and errands. You decide to take a look at your recent data to get a sense of anything that needs taking care of.”

Scenario 4 (protecting data privacy): “You have heard about a data leak from a popular cloud storage platform that exposed personal information to hackers. You want to review what data you have stored on different cloud platforms that might pose a privacy risk in the future.”

Scenario 5 (safeguarding data across devices and platforms): “You are in the process of buying a new computer. You want to make sure that you are not going to lose any of the data you care about. You want to ensure that everything is stored in more than one place.”

The scenarios are based on insights from previous chapters: they consider both different decluttering strategies (Chapter 5) and different behavioral styles in personal data curation (Chapter 6). We wanted to explore how different functions in the prototype can support different scenarios and user attitudes.

More specifically, Scenario 1 (space running out) closely resembles the context behind *triggered* decluttering (Chapter 5) and reflects goals common to all behavioral styles, with the Overwhelmed style being a stronger target than others (Chapter 6). Scenario 2 (regular data curation) ties back to *routine* decluttering (Chapter 5) and largely reflects goals from the Purger or Frugal behavioral style (Chapter 6). Scenario 3 (exploring recent data) is mainly targeted towards the Casual behavioral style (Chapter 6) and also reflects a *serendipitous* decluttering strategy (Chapter 5). Scenario 4 (protecting data privacy), focuses on the cross-behavioral style goal of protecting data privacy. Finally, Scenario 5 (safeguarding data across devices and platforms) reflects one of key goals for the Collector behavioral style (Chapter 6).

The scenarios were also key for helping participants focus on concrete implications of use rather than low-level details of the prototype (e.g., colors, fonts, buttons). Some of the scenarios mention specific devices to feel concrete, but we encouraged participants to see them as a starting point to discuss additional devices and broader situations or practices.

Debriefing interview

The third and final part of the session was a debriefing interview about the prototype. Here, we asked participants about their impressions of the system, clarifications about what they did during the scenarios, and then walked through all the four sections of the prototype one by one to gather more specific feedback. Finally, we had a short card-sorting activity where we asked participants to rank the five scenarios by how relevant they were to their own experience. We prompted participants to explain their ranking, elaborate on the match between different data management methods or tools (prototype included) and scenarios, and consider in what other situations they could imagine using the prototype. Participants used small paper printouts of the scenarios. We also asked participants to rank the usefulness of the prototype against any tools they mentioned using in the screening survey or during the interview. Once again, participants used paper printouts with the names of the different tools we had prepared for them.

8.4.3 Data analysis

We used thematic analysis to develop recurring themes and patterns from the sessions [61]. The analysis process took place over three weeks. Two members of the team conducted the bulk of the analysis and later discussed the themes with the other team members. We started with a round of open-coding, where the two team members coded data in parallel, seeing each other's codes. Then, we grouped codes into categories, and started thinking about themes and patterns across categories. Codes and categories were both inductive and deductive (based on insights from previous studies, and specific aspects or sections of the prototype). We discussed several iterations of possible themes, choosing specific areas of the analysis to focus on. After identifying the key lens of the analysis (centred around data boundaries and their effect on centralization and customization), we went back to the transcripts and re-coded them using only the three themes to check for consistency and make sure that our interpretation fully captured participants' experience.

8.5 Results

In this section, we first provide some context for the analysis, based on general reactions to the prototype and how participants used it during the scenarios (Figure 8.7) (more details are also available in Appendix A). Then, we delve into the more interpretive part of the analysis. As in Chapter 7, the themes we present are a high-level cross-cutting synthesis of individual behaviors.

8.5.1 Contextualising information

Overall impressions

Most participants (12/18) had positive reactions to Data Dashboard, saying that it was "smart," "intuitive", "user-friendly," and would save their time. Several participants also preferred the system when comparing it to other tools they had used in the past. Some participants went so far as asking if they could have the system installed on their devices after the interview.

But not all participants liked Data Dashboard. Participants who had mixed reactions (3) thought that some aspects of the system were unclear or unnecessary.

Participants who had negative reactions (3), instead, said they did not need or want a tool such as Data Dashboard. Some opposed the idea of a system deciding how to curate data. Others did not see data management as something worth their time. These negative reactions are in line with previous work and support the idea that different users have different needs when it comes to personal data curation and technology support, as shown in previous chapters. Although it was not our goal to find precise correlations among behavioral styles and reactions (the number of participants would not allow it in any case), we noticed that the negative reactions came most commonly from participants who took a collector approach–this result aligns with our expectations but future studies with a bigger sample can better explore this possible link.

Many participants also reflected on the potential privacy risks of centralizing personal data; we expand on this theme later.

Interaction during the scenarios

Explore Your Data (EYD) and Quick Actions (QA) were the most commonly used sections of the prototype during the scenarios. Participants thought that EYD worked best for occasional, more focused scenarios and they used it more frequently in scenario 1 (getting rid of data to free up space) and scenario 2 (regular scheduled cleanup). Instead, QA would work better for short management episodes (scenario 3, taking a few minutes to look at recent data). Most participants also used Activity at some point, but several participants found it underwhelming and too similar to EYD. Several participants did not notice the Shared Data section or found it confusing. Some thought it would help them when working on collaborative projects.

All participants except one discovered the sidebar filters in EYD and QA on their own, and most used them at some point during the scenarios. Most participants found the filters comprehensive and the idea of automatically clustering data helpful. Participants did not have clear requirements for how the system should generate suggestions or cluster data, but they expected "*a machine learning algorithm that gets better with time.*" (P18) Most participants also discovered the Settings page and its link to the sidebar filters but several found it initially confus-

ID	General reaction	Most used section	Most relevant scenario	Least relevant scenario
P3	positive	Explore	(5) Storing on multiple places	(4) Privacy leak
P4	negative	Explore	(3) Exploring recent data	(4) Privacy leak
P5	more negative	Explore, Quick Actions	(2) Regular cleanup	(3) Exploring recent data
P6	positive	Explore, Quick Actions	(1) Space running out	(4) Privacy leak
P7	positive	Quick Actions	(3) Exploring recent data	(4) Privacy leak
P8	positive, mixed	Quick Actions	(5) Storing on multiple places	(4) Privacy leak
P9	more negative	Quick Actions	(1) Space running out	(5) Storing on multiple places
P10	mixed, more positive	mix	(3) Exploring recent data	(4) Privacy leak
P11	positive	Quick Actions	(3) Exploring recent data	(4) Privacy leak
P12	positive	Quick Actions, Settings	(3) Exploring recent data	(5) Storing on multiple places
P13	positive	Explore, Quick Actions	(1) Space running out	(3) Exploring recent data
P14	positive	mix	(1) Space running out	(4) Privacy leak
P15	positive	Explore	(2) Regular cleanup	(1) Space running out
P16	positive	Activity	(2) Regular cleanup	(5) Storing on multiple places
P17	positive	Quick Actions, Explore	(1) Space running out	(3) Exploring recent data
P18	positive	Explore, Quick Actions	(1) Space running out	(3) Exploring recent data
P19	positive	Quick Actions	(2) Regular cleanup	(1) Space running out
P20	mixed	Explore	(3) Exploring recent data	(2) Regular cleanup

Figure 8.7: An overview of participants' varied reactions to the Data Dashboard prototype. For each participant, we report the general reaction (negative, more negative than positive, mixed, more positive than negative, positive), the most used sections during the scenarios we provided in the evaluation, and the participant's ranking of the most and least useful scenario. As noted earlier, we excluded P1 and P2 from the final analysis.

ing. In general, participants thought that it was a good idea to customize the filters because it gave them more control; we expand on this theme later.

Scenario 4 (protecting privacy) and scenario 5 (safeguarding data across devices) presented some challenges for participants. When going through the privacy scenario, several participants scanned the system looking for a way to see all data stored in a specific platform (e.g., iCloud, Dropbox, Google Drive) or device (e.g., "my drive"). Often they could not find what they were looking for. But many participants also said that they prefer not to store private data on the cloud in the first place and that this scenario did not apply to them. In the last scenario, instead, many participants did not use the system and talked about their actual process of backing up files either manually or through completely automatic solutions (e.g., Apple's Time Machine, Google Photos). Some participants also had trouble with the language used by the system, saying they had no idea what "syncing" data meant: "*I know what data to back up means but I don't know what sync means. I don't know if they're related, maybe.*" (P16) Overall, these reactions highlight how participants saw privacy protection and backing up as processes that either take place outside of specific tools or require little input from them. Participants most commonly ranked these last two scenarios as the least relevant.

These results show how our scenario-based evaluation prompted participants to explore Data Dashboard and think about its design. They also suggest that our approach of combining centralization and customization has potential. To unpack its value and potential risks we now turn to the more interpretive part of the analysis, where we delve into key aspects of curation, centralization, and customization.

8.5.2 Data boundaries drive curation

When we analyzed participants' reactions and interactions with the prototype, one idea became key: data boundaries drive curation. We have briefly discussed what we mean by data boundaries in the related work section. Boundaries are an abstract concept that can explain how people enact curation of their personal data ecosystems [288]. People create implicit or explicit boundaries that separate different categories of information. These boundaries help people build their identity, mark areas of their life, and feel in control. For some people boundaries will be more malleable, for others more strict [219]. Here, we provide additional evidence for how people create, protect, and think of data boundaries. Then, we use the concept of boundaries to explain reactions to centralization and customization.

Creating boundaries

All participants implicitly talked about creating boundaries when deciding what data to store and where to store it. They mentioned rules for what goes where and why: "*I have my data compartmentalized:* [the] *tablet is just for reading.* [The] *laptop is for everything school related.*" (P9). A key boundary was between private and non-private information, with participants choosing where to store private information based on perceived privacy risks of different devices and cloud platforms: "*I don't put any of my private stuff on the computer at all.*" (P6) Boundaries also helped participants distinguish and prioritize data based on importance. For example, P10 (who showed a mix of collector, purger, and overwhelmed behaviors) discussed moving data from one cloud platform to another to separate expendable and essential information: "*I started moving everything* [from Box] *to Dropbox, because I think it's just more reliable. So everything that I can't afford to lose, I'll stick it in Dropbox.*"

Breaking boundaries

The centralized approach of Data Dashboard prompted participants to reflect on how they often experienced breaking points in boundaries. Sometimes this was intentional. Reflecting on the suggestions and filters for "private information" in the prototype, one participant explained that sometimes it is necessary to break a boundary around privacy because of convenience: "[Having passport information on Google Drive is] *not ideal. But because of frequent travel, I am somewhere and I am filling in a form and I need that information,* [so] *either through Google Drive or email I was able to locate it. So, it's more about ease of access to the information that saves my time than anything else. Ideally, I don't like to have important documents even on email, but I haven't learned if there is a secure system to store them online.*" (P17)

In other cases, boundaries were broken unintentionally. For example, after looking at the different types of data in the prototype, one participant explained the frustration of having WhatsApp photos from other people go automatically into their own. In this case, the boundary between "my stuff" and "other people's stuff" was broken without permission: "*I tried to figure out on this device as well, some sort of filter where you can manage whether your images from WhatsApp, a group message thread, are automatically going into your photos or not. It drives me crazy that WhatsApp's photos go automatically into your photos. If there's more of a filtering system to help you organize that, that would be great.*" (P12)

Boundaries influence trust

The need to create and protect boundaries also influenced participants' trust in the systems and platforms to manage data. Several participants asked whether Data Dashboard would be associated with a specific brand because they tended to trust specific brands with their data: "*Google has created credibility over years so I know I have trust in that system.* [I use] *Mac because of* [its] *ease of use but I don't always store all my important info over there. They have a very tricky way. If you're locked out, only they can unlock it. I don't like that dependency on a third party.*" (P15) Others, instead, referred to the boundary between physical hardware and cloud platforms to explain their practices, wondering whether Data Dashboard would be largely local or cloud-based: "*You can't trust anything that you don't have control on the hardware. Especially with smartphones, because most of it goes through the cloud, you can't really trust anything,*" said P4, who had both a collector and frugal approach with data.

Boundaries lead to fragmentation

A consequence of creating boundaries is data fragmentation [143, 288]. This is one of the key challenges for data curation in an age of multiple devices and cloud platforms. When considering data boundaries, participants explained how fragmentation could be beneficial: "*I am isolating my data more and more rather than sharing it. The people who invented Facebook and all these platforms, they did it with good intentions. The problem is these platforms are now abused* [...] *You can't really trust their intentions so you have to protect yourself by isolating your data.*" (P4) But it could also be costly, leading to confusion and frustration: "*I've started working for different organizations, you have data from different places and sometimes it gets confusing. I have some things from my previous experiences and now everything is getting mixed up.*" (P3) From a design standpoint, we can ask whether fragmentation is a problem to be solved and how to solve it if that is the case. Next, we look at how centralization and customization intersect with data boundaries.

8.5.3 Centralization blurs data boundaries

Centralization is convenient

Participants saw something positive in centralizing data, with the different sections of Data Dashboard providing avenues to reconcile data from different devices and platforms: "*I really like how it's combining different sources of where these things could be stored.*" (P12) They thought that centralizing data provided some clear benefits, like saving time: "*This looks delightful. The most important criteria for a service like this for me is the time benefit I get from it* [...] *I don't mind giving them access to all my data.*" (P3) Or, making it easy to see everything in one place, as was the case in Explore Your Data: "*You see everything: your contacts, your bookmarks. That's really handy. I like that. It just feels very comprehensive.*" (P10) "*It's all different types of* [accounts]. *I like the way it's set up. It's quite clear, it looks good. I can see everything at a glance.*" (P16) They also hoped that the cards for different types of data they saw in Data Dashboard could break the barriers between different devices: "*If it were to show me text messages, it would be good for me because right now I have no way to see them from the computer.*" (P11)

Convenience goes hand in hand with risks

However, data centralization introduced a dilemma. As much as it seemed convenient, it created additional risks: "*If I have everything together,* [the] *ease of use is high but* [the potential risk for a] *security breach is higher as well.*" (P15) Especially when looking at Activity and Explore Your Data, most participants expressed concerns that touched on how centralization blurs data boundaries: "*Some people like to collaborate in one place, one app, access* [all] *email accounts and whatever, but that's not my preferred way.*" (P8) Based on these concerns, they wanted Data Dashboard to respect the boundaries that they created by choosing specific storing places for different types of data (a result consistent with past work [166]): "*I already upload most of my files to Google Drive, Dropbox. I don't feel that bothered. But if this accesses even the files that I specify not to be* [accessed for a specific function], *that would really bother me.*" (P5, whose approach to curation was a mix of collector and casual.)

In particular, when considering privacy data boundaries, several participants felt uneasy about Data Dashboard: "*You already have questions about the security of iCloud, Google Drive, etc. and that's reduced when you go to third-party tools. Especially when you see* [that] *the algorithm can see private information.*" (P18) Some participants imagined the negative consequences of having all data centralized in one system and wondered whether this is the right approach. For example, after seeing a recommendation to review cloud documents containing passport or credit numbers in Quick Actions, one participant wondered about the consequences for privacy: "*But there is a link to the file and it says what's in the file, so what else do I need as a thief? You just gave everything on a platter. It doesn't make any sense to me.*" (P4)

These reactions resonate with the *privacy-utility trade-off*, a common notion across past work on smart energy systems [282], health data [48], location data [52], and differential privacy [6] among others. In short, the privacy-utility trade-off considers the tension users might experience with a program that uses data about them to provide a potentially useful function (e.g., recommending what movies to watch based on movies watched in the past).

Some participants also had issues with terms related to privacy in EYD, saying they were unsure of what "tracking data" meant: "*Tracking data.... what does it include?.... Not totally sure.*" (P12) We used these terms because they are common in similar tools, but these reactions show the importance of language for building trust in the system and building awareness of what private data users might have stored on their devices.

Centralization requires guarantees

Because of the potential to blur and distort boundaries, participants expressed a need for explicit guarantees around centralization: "*I just want it to be secure* [so that] *no one else can get into the system. I don't know how... I have no idea how you do it. How do you protect something like that?*" asked P16, one of the participants who felt overwhelmed with data. Participants wanted to make sure that the system would respect data boundaries and mitigate potential security risks: "*It's the central point that commands all my accounts so it has to be highly secure.*" (P7)

Participants rationalized the potential adoption of Data Dashboard by referring to the terms and conditions that the system would have and would need to strictly apply (although there was no such thing in the prototype): "*If I am using the system, the terms and conditions that are mandatory to be agreed upon, should not include that your data has any potential of being shared by any third party for any commercial use.*" (P17) If the terms are clear and the system feels reliable, then centralization becomes acceptable: "*It depends on the privacy statement really. If the agreement seems good enough and it seems a reliable service, I don't mind using it and it going through my things.*" (P3) If it is possible to reinforce a boundary between local data and data in the cloud, then the worries disappear: "*If this is an online interface, I'd have problems with it. But if it was offline with no interaction other than backing up, which I control, then I'm completely fine.*" (P5) Once again, these reactions highlighted the need for clear explanations around data curation, with participants often blaming themselves for being "not techy".

Even with potential issues mitigated, some participants wondered about the feasibility of centralization. They reflected on the underlying conflict between their expectations and business practices that impose borders around data: "*My question is, can you actually make it? And I'll be the first guinea pig. The big giants, they have big muscle and they compete.* [...] *I don't know if it's possible because of the conflict of interest of these business things.* [...] *I am using Google Drive because I have a Google account, iCloud* [because I have] *an Apple account. Does this* [system] *need to be linked to a certain company or email account or something? If this is like my online version of my external hard drive, yeah, I would like to have everything consolidated under the one big, safe roof.*" (P19) This conflict required a broader guarantee, one that puts user needs above business needs: "*If there is such a magnificent creation down the road, it could be under the one roof, but also secure and safe. That's kind of my dream.*" (P19)

8.5.4 Customization upholds boundaries

Customization can make boundaries visible

Most participants had positive reactions to the customization options in Data Dashboard: "*I like choices, so I'm all for it.*" (P17) Some were confused by the link between the sidebar filters and the Settings page, but in general they thought that customizing was a good idea. For example, participants liked the option of sorting and hiding data types in Explore Your Data because it helped them see the things they wanted to prioritize: "*I'm the sort of person who would limit what I see here. I would keep photos, bookmarks* [...] *It's also good that I can reorganize. Minimize what I see and prioritize what I see first.*" (P9) Similarly, the option of sorting through data using the personalized filters excited participants because it made it easier to look at their data: "*This is great to look at what's old, what's inactive, unreachable, unused, these are really good,*" said P12, who also described herself as "a purger more than a saver, always looking for ways to free up space." Together, these options made personal boundaries more visible and tangible, and helped participants navigate different data curation scenarios: "*Sometimes I need to see all of the information, sometimes if I am running on short time, I wanna do a narrow search, instead of always seeing everything.*" (P17)

Customization allows users to manipulate boundaries

A consequence of perceiving boundaries as more visible was that participants also felt they had more options for manipulating them. Several participants thought that the default options in the system were "novel" and "comprehensive" enough to meet their needs. Choosing what to include or not made them feel more in charge of the system: "*I have some kind of authority to make a selection of what I want. This is my priority, I want to check duplicate documents. These options are good because then I also have a sort of selection power.*" (P15) But an additional positive aspect was the option of creating your own filters through custom combinations in Settings. While this function was confusing for some participants, they generally saw it as a useful way of setting specific boundaries: "[It is] *cool* [that] *you can* *make your own filters. I like that 'cause I can figure out what's important to me, my own criteria, I like the customization aspect.*" (P18)

Customization reinforces control

One possible disadvantage of customization is the time required for it, as we mentioned in the related work section. When discussing the customization options in Data Dashboard, and the Settings page in particular, participants reflected on the time required for setting the filters as they preferred. Several participants imagined it would not take a lot of time because the system provided default options to choose from: "*I don't think it would take that much time.*" (P12) Others imagined it would take some planning on their part: "*I think it might take me probably a couple of hours to figure this out. I would plan it out before I actually use the system.*" (P10) But in the end they thought that investing time in customizing was worth it: "*I don't mind investing the time if this was going to figure out the system, I would get out a piece of paper and think of recent data to get rid of.*" (P10). Taking the time to customize would lead to the system respecting their own priorities and boundaries: "*I think filters are always useful. I don't mind* [spending time customizing] *because everyone has something different they're looking for so it's good to have these filter so they have what's most important to them.*" (P3) Participants perceived this trade-off between time and control as necessary to counterbalance any potential risks coming from other aspects of the system. Feeling in control made them feel safer: "*When it comes to data, I am very cautious. So if I have the settings set correctly, hopefully it acts accordingly.*" (P19)

Overall, these positive reactions show that customization upholds data boundaries by making them more visible, allowing users to manipulate them, and reinforcing their control and involvement in data curation.

8.6 Discussion

Our results around centralization and customization show the importance of data boundaries for curation. While the centralized nature of Data Dashboard had several advantages for curation, it also had the potential for blurring boundaries. Past work on centralized personal archives [166] similarly found that storing personal data in a central place undermines the "different facets of the self," ignoring differences between unremarkable and valuable content. Our study suggests that customization can offset the negative aspects of centralization and better help users demarcate different types of content. Customization options, on the other hand, tended to uphold participants' boundaries. Combining the two approaches is a promising design direction that can balance conflicting user needs. Below, we reflect on some of the evaluation results and outline how to move forward in designing curation tools.

8.6.1 Reflecting on the feasibility of Data Dashboard

Before discussing potential directions inspired by our results, we address a key question: as P19 wondered, is a system such as Data Dashboard feasible? Would limiting its access to metadata be enough to make it viable? Would "the big giants, with their big muscle" allow it? Probably not, because each of them is making the same promise to users: enjoy simplicity by having all your data in one place! Our place. Except, as a user you might end up with several places, competing with each other [288]. At the moment, there is little benefit for cloud storage providers to reduce fragmentation and provide data interoperability, but things might change in the future. For example, Dropbox has recently started to allow its users to create, open or edit Google Docs files from within the Dropbox interface (but without the permission to move files off Google's servers) [80]. In turn, Google Drive allows users to edit Microsoft Word, Excel, and PowerPoint documents. In theory, these are steps in the right direction, but similar "integrations" still rely on proprietary formats and depend on the whims of a small set of companies. Is there another way? Yes and there is a precedent: email standard protocols.

IMAP, SMTP, and POP3 email protocols allow users to access and manipulate emails from a variety of clients, often leaving data on the original cloud server where it is stored. Users can also aggregate multiple accounts into a single client, easing the burden of fragmentation while still enforcing personal boundaries. Similar standards should apply to a much broader set of data types outside email. The continued introduction of new email clients year after year also proves that standards do not limit commercial opportunities. Once access and basic actions become standardized, tool creators can focus on innovative functions that address more interesting, unexplored user needs and make their product unique. Below, we discuss some ideas for future developments.

8.6.2 Integrating data boundaries into design

An implication from our work is that filtering and sorting data into "chunks" based on types or other automatic categorizations can be helpful. The examples in Data Dashboard suggest that working on algorithms that can filter and recognize different types of information is a promising direction. While there are previous technical efforts along these lines [16, 64, 261], more work is necessary to address the functional design of these mechanisms.

Current commercial products for storing and managing personal data also offer some form of support for enacting boundaries, but often in a basic way and with limited scope. For example, cloud storage platforms such as Google Drive, Dropbox, or OneDrive allow users to set sharing settings for documents. But this type of boundary mechanism is often static, only possible at the level of individual items, and tends to focus on privacy in the context of sharing data with other people. Instead, we have seen that data boundaries are dynamic and contextual, not only about privacy, and most often involve broad mental categories. Similar mechanisms are also often inconsistent across services and require users to set similar settings multiple times in different storing places. This limitation highlights the potential value of a system such as Data Dashboard.

Another example of boundary mechanisms comes from email clients such as Gmail, that allow users to automatically sort emails into different categories (e.g., social, promotions, updates) (Figure 8.8). This is a promising idea. However these categories are pre-defined and users have little freedom in personalizing them. Overall, we think that support for data boundaries could be more dynamic, expressive, and cohesive.

One possibility is to envision more granular mechanisms that integrate data boundaries in their design and combine the key positive aspects of both centralization and customization. For example, we can imagine users being able to directly review and manipulate boundaries at the level of individual items: they could be

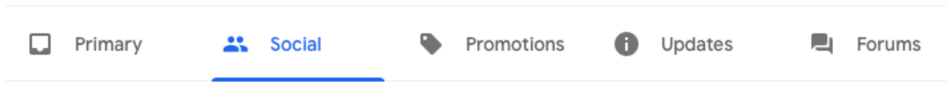


Figure 8.8: Gmail allows users to automatically sort emails into pre-defined categories.

looking at a document containing their passport number, and choose how private, important, or relevant they consider that type of information to be. Then, they could set the boundary to be valid only for a certain amount of time and apply it to similar items or just the single item, set it for one platform or multiple, all from a central control point. An explicit boundary management mechanism could allow users to better manage breaking points, defining whether a boundary can be broken and in which scenarios (e.g., "keep passport data off the cloud, unless I am travelling").

Another potential avenue for integrating boundaries into design is to think of them as objects that users can share and exchange. The initial premise of exploring customization was to see if it would help in the design of a system that can accommodate individual user needs. Our results show that this is possible. However, while each person is different, there might be clusters of users who take a similar approach to data curation. We can envision a way for users to adopt the boundaries of other people, adapting them to suit their own needs. A system similar to Data Dashboard could use a similar sharing infrastructure to provide more personalized defaults based on user preferences. This could reduce the time required for setting personal filters and make the system more appealing to those who are not willing to invest time in customization [120]. Of course, it would be necessary to consider how to create similar functions in a privacy-preserving way, so that they do not break the boundaries they are trying to support.

8.6.3 Rethinking the language of personal data

Another key line of work emerging from our study is about rethinking the language of personal data management and curation. Personal data curation has relevance beyond practical actions such as deleting files, moving documents into folders, or uninstalling mobile applications to free up storage space. Those are concrete actions that need support. But the broader implications of curating data are about control over personal information and the consequences of producing, storing, and accessing data in platforms and devices that are part of a "surveillance" apparatus [319].

The language used to explain where personal data (from documents to location history) is stored and what happens to it is essential for understanding and controlling its use. In our study, we saw that participants sometimes struggled with some of the terms used in the system, such as "syncing data" or "tracking data" and pointed to the importance of language and terms for building trust in the system. These recurring reactions from a varied sample of "not techy" participants show how it is necessary to work on improving the language of everyday user interfaces, especially in privacy and security-related scenarios. How can users curate and control their personal data if they are not aware of what this category might include and what common actions in everyday tools do? Past work shows how language inconsistencies are common in popular operating systems for actions as simple as deleting data [113]. We argue that simplifying and unifying the language of personal data is a necessary, basic way of supporting user needs. Doing so would be the first step to address the deeper issue of naming the processes behind personal data in a way that makes people more aware of the mechanisms behind its storage and handling. As Zuboff says in describing strategies for fighting surveillance capitalism, naming is the first step in "confronting and taming the unprecedented." [203]

To help navigate a looming data-driven society [86], privacy and data regulations are paramount [1, 265], but design practice has a role to play too. There is an opportunity for future design initiatives to start addressing the underlying gap in personal data literacy. This would involve rethinking common technical terms for describing personal data, and creating initiatives, both within tools but also outside them, to build a common vocabulary around personal data. We imagine this as a community effort where designers and researchers explore how user language and the grammar of possible actions within data management are intertwined, an idea brought forward by previous work [123]. One option could be to promote consistency across operating systems, applications, and tools, creating a grounded, standard vocabulary that users are familiar with. This could involve a systematic study of users' term that could be then integrated into "personal data standards," just as there are conventions for common design elements. Another option would be to let users define and teach their own language to the tools through mechanisms that leverage a link between individual items and categories of data. For example, systems could give users an example item, ask them to categorize it using their own words, and then apply this user-generated vocabulary in the interface.

8.6.4 Centralization as a matter of perspective

Another important thread in our work is the contrast between physical devices and cloud platforms, with centralization hinting at a tension between local and online data storage. This tension highlights a gap between users and technology companies, with different perceptions of what centralization means. For users, centralization is largely about data *access*. For companies, centralization is instead about data *storage*.

Users tend to perceive local devices as closer to them: after all, personal devices used to be the only place where users could access their personal data. Cloud platforms, instead, feel more peripheral: they have become a key component of personal data ecosystems [288], but because users tend to use more than one, they can feel fragmented, less tangible, and out of sight [213]. Thus, for users, accessing data from the cloud is about bringing items back from a set of far away places to a central device in their hands. This is why the Cloud, to users, can feel decentralized. But from a technical standpoint, cloud computing is in fact a centralizing force [94], pushing all data to be stored not on single devices but on the servers of a few large companies, centralized in a limited number of locations. These companies then promote their cloud platforms by emphasizing the convenience of storing everything into a central place. Except, people's perception of what is central is different. This divergence in the promised vs. the actual experience of the Cloud reinforces the need for more consistent language around personal data and stronger conventions. It also makes us wonder whether a different paradigm for modern computing is possible.

8.6.5 Envisioning a post-cloud future

Current setups force users to go through a third party for accessing data across devices. There is no practical alternative. This dynamic has consequences for personal data boundaries, with tools creating boundaries of their own and imposing specific structures on data (e.g., having to store files in a Dropbox folder if you want to access them on more than one device, rather than in a local and more personal structure). But what if we imagined an alternative that emphasizes users' perception of local devices as central, rather than giving priority to the cloud? For example, one where interoperability across data formats and storage devices made it possible to easily synchronize data without the need for a cloud-first structure. The tools and user interfaces would be similar to existing ones, but the functional paradigm would be different. We imagine that future data infrastructures could rely on local, "social" clouds, maybe with a physical device acting as the central node connecting data links across devices. For example, members of a family, neighborhood, or social group could set up a "local cloud" that allows personal data to move freely among known and trusted members' devices. Items could be synchronized to different devices on an individual basis rather than by uploading them on a distant server. The local cloud would take advantage of the multiple devices for maximizing storage and preventing data loss through redundancy. A similar approach could make it possible to make data available across devices, while still respecting data boundaries and local structures. These ideas are speculative, but not entirely new. Before the advent of commercial cloud computing, investigations on peer-to-peer and decentralized technologies for data access and storage were popular [3, 71, 146, 244]. But then, the Cloud won. Today, the increasingly critical discourse around technology companies, cloud platforms, and privacy might signal a shift in perceptions. Within this evolving landscape, there is renewed interest around decentralized computing [297], cooperative storage [275], and cross-media information spaces [257]. All together, current efforts and past explorations point to a promising domain that future design work can help explore [287].

8.7 Limitations

Our sample is meant to be generative and not statistically representative, so it has some limitations around age and gender balance. We did not screen for gender identity, but studies on related topics have similar samples [58, 59] and notice no apparent differences in participants' attitudes based on gender. We screened for age, but we did not get interest from many respondents over 50. Instead, we focused on the variation in data curation approaches. Recent work [5] hints at differences in desired PIM behaviors based on age and gender, providing a complementary perspective to our work. Future studies can further complement our work with a more representative sample to see if and how our results transfer to a broader population. Future work should also explore quantitative measures of user satisfaction for a design approach similar to ours, how to extend our efforts to other stages of data curation (i.e., retrieval), and what might be the psychological or logistical effects of spending time curating data (e.g., enjoyment, satisfaction, time saved, more efficient retrieval).

8.8 Summary and conclusion

Storing, managing, and curating personal data is a challenging process made more complex by the mix of devices and cloud platforms that people regularly use. In this chapter, we have explored how centralization and customization can support people's behaviors. We identified how centralization blurs personal data boundaries, while customization upholds them. Using this specific analysis lens helped us outline key challenges for designing data curation tools. As our relationship with data evolves and likely becomes more complicated, we see taking a step back and reflecting on key design decisions as essential for shaping future products. Our work shows that there is great potential for exploring how to design tools that integrate data boundaries as a core part of their functionality and provide new mechanisms for managing them. We hope our work will provide a foundational starting point for innovative tools that can define future paradigms for data management.

Chapter 9

Conclusion

In this dissertation, we have presented the results from four interview studies (total of 64 interviews) and an online survey (349 respondents) on personal data curation behaviors. We have investigated user decisions around what data to keep or discard and characterized individual differences in this space. Starting from an initial spectrum of general tendencies for preserving data (Chapter 4), we have identified five behavioral styles that describe contextual patterns in participants' curation approaches (Chapter 6), making individual differences actionable for design. Then, we have used these insights to explore alternative design concepts for selecting data (Chapter 7) and a personalized approach to personal data curation (Chapter 8).

Below, we first summarise these and a few additional contributions in more detail. Then, we discuss implications for future work. We conclude with a few closing remarks to reflect on our work.

9.1 Summary of results and contributions

In this section, we summarize the key results and contributions of the dissertation, distinguishing between primary and secondary contributions.

9.1.1 Primary research contributions

Rich characterization of data preservation tendencies

Before our investigation, the HCI literature lacked a deep, systematic understanding of how people decide what data to keep or discard. Most of past work relies on the assumption that people tend to just keep everything. But that was a broad assumption and in this dissertation we show that there can be substantial individual differences in how users decide what personal data to keep or discard.

Our first key contribution is a rich characterization of general data preservation tendencies and their role for identity construction.

In Chapter 4, using 23 interviews with a broad sample, we identify a spectrum of preservation behaviors with two extremes: "hoarding" on one side (where participants tended to keep most of their data over time) and "minimalism" on the other side (where participants tried to keep as little as possible). We show how these behaviors are contextual, nuanced, and important for identity construction.

Previous studies on personal data and PIM also highlight the tie between data and identity, but our work shows how this connection can be experienced with opposing attitudes ("I am my data" vs. "I am more than my data"). This distinction is subtle but key to understanding how different people might approach data curation using different strategies and approaches.

Five behavioral styles that synthesize personal data curation approaches

Building on top of the "hoarding" and "minimalism" spectrum, we identify a set of five behavioral styles and approaches to personal data curation, using a total of 64 interviews from the four main interview studies that make up the dissertation (Chapter 6).

The five behavioral styles (*Casual*, *Overwhelmed*, *Collector*, *Purger*, *Frugal*) complement previous characterizations of individual differences in PIM research with a focus on keeping and discarding decisions. They also represent a significant advancement of our initial characterization of user behaviors in Chapter 4. They enrich our understanding of personal data curation as a practice tied to identity curation, bringing to light a temporal dimension to the process.

The behavioral styles also represent an actionable and generative resource for design that practitioners and researchers can use.

Design dimensions and design concepts for supporting data selection

In Chapter 7 we bring together insights from previous studies and past work on digital data to explore a design space of possible solutions. We describe four key design dimensions that can be a helpful resource to generate design ideas: *selection regime*, *automation*, *system aggressiveness*, and *temporality*.

Then, we use the dimensions to create a set of five design concepts that map the design space and explore alternative approaches for supporting data selection across a range of situations and devices (*Patina*, *Data Recommender*, *Temporary Folder*, *Temporary App*, *Future Filters*).

These speculative concepts can be a starting point for more comprehensive products, whether in industry or academic research.

Our methodological approach of creating video prototypes for the concepts can also serve as inspiration for future studies that explore uncharted design spaces where technical constraints limit possible prototypes.

Example of a personalized approach to personal data curation

Our work in Chapter 8 combines insights from previous studies into a cohesive system, *Data Dashboard*, that focuses on exploring and evaluating two key design approaches for personal data curation: centralization and customization.

Our results show that this approach can be successful in supporting personal data curation behaviors. However, it needs to respect and integrate data boundaries (permeable, invisible rules that inform where and how people store personal data) as a first-class object for users to interact with.

These insights provide a concrete starting point for building new products and advancing user support for data curation.

Generative design directions based on empirical work

Throughout the dissertation, we propose several generative design directions and opportunities that can help in prioritizing user support, exploring new product concepts, and expand the focus of personal data management tools.

We see these generative design directions as broader, more open-ended, and less prescriptive than traditional design implications found in many HCI studies. The directions we propose indicate open opportunities but do not necessarily prescribe how to implement them [78, 248].

At a high level, we propose to tailor technology to different tendencies for data preservation, finding ways to offset costs inherent at both ends of the hoarding-minimalism spectrum (Chapter 4). We also highlight design opportunities for supporting data decluttering by visualizing temporal aspects of data, proactively recommending items to discard, and preventing unwanted data accumulation (Chapter 5). Then, we outline ways for prioritizing user support, making curation feel more engaging, exploring the role of curation for memory, and investigating alternative ways of capturing user variation based on the five behavioral styles we present in Chapter 6. We follow with directions for centering design work around personalization, automation, and privacy by exploring safeguarding mechanisms and new data curation actions (Chapter 7). Finally, in Chapter 8 we discuss ways for integrating data boundaries into design, rethinking the language of personal data, and envisioning a post-cloud future.

These directions can be a starting point for future research and design agendas, paving the way for innovative data management tools that will help shape the upcoming post-cloud landscape (Chapter 8).

9.1.2 Secondary research contributions

Reflexive account of our user modeling process

Our research process in Chapter 6 can inform similar user modeling efforts in other domains. The reflexive account of how the behavioral styles evolved throughout the four studies, and how different methods helped us enrich them, can help researchers or practitioners to carry out a similar modeling process for capturing salient individual differences.

Taxonomy of personal data types and decluttering criteria

In Chapter 5 we build a taxonomy of data types and related decluttering criteria, using an online survey with 349 respondents. We identify six macro-categories of data types (Documents, Organization, Communication, Media, System data, Logging data) and map different decluttering criteria to each of the categories.

These categories and criteria can be used for designing new tools, scoping research studies, or directing algorithm design.

Description of temporal decluttering practices

Finally, still in Chapter 5, we identify a set of temporal decluttering practices (*routine*, *serendipitous*, *triggered*) that can inform design and provide support for the behavioral styles and approaches we present in Chapter 6.

9.2 Implications and future work

Next, we outline implications for future work, touching on how our results can be leveraged for User Interface (UI) and product design, HCI research, algorithm design, policy and sustainability.

9.2.1 Implications for UI and product design

A key direction for future work is to build upon our efforts on differentiating, prioritizing, and personalizing user support. In particular, we encourage future design work to explore alternative ways of addressing personalization. For example, the Settings in Data Dashboard provide a first step in this direction, but future work should test and evaluate different visual approaches for providing customization options that go beyond a set of check boxes.

Similarly, a key implication of our work is to keep exploring and pushing new design metaphors for conceptualizing personal data. For example, exploring a positive take on curation by bubbling up data at the right moment and building sediment into personal items. Or, exploring how concepts such as fossilization and disintegration can apply to personal data. Bubbling data could involve shifting the focus from active curation to passive curation, with systems reducing user effort by resurfacing relevant items based on content similarity or context-matching (e.g., a photo of an event bubbling up in a message conversation when the event is being discussed). Metaphors around sediment, fossilization, and disintegration, instead, could focus on expanding dynamic and temporal qualities of data from a visual standpoint. As an example, a segment of data could become impossible to edit after a set amount of time to visually signify its outdated relevance, thus fossilizing. Similar metaphors provide a starting point for exciting new interfaces that move beyond current abstractions, largely tied to office work (as is the case with files and folders) or technical implementations (as is the case with mobile applications).

These ideas, together with the design directions we propose throughout the dissertation, point to the need for rethinking personal data management tools with a long-term perspective.

9.2.2 Implications for future HCI research

From a research perspective, the key implication for future HCI work is to expand the focus of our investigation in terms of users, data types, and methods.

Our work focuses on personal data curation for individuals. However, a natural extension of our work would be to look at individual differences in curation in a collaborative setting, as we mention in Chapter 4. Future studies can use our work to investigate how different behavioral styles and approaches influence collaboration, both in work and domestic settings. For example, how do individual differences in curation play a role in small, unstructured groups (e.g., freelance knowledge workers, student groups, and so on)? How can collaborative tools better accommodate different management and curation styles? How do similar insights shift when considering shared data in a family context? And what are the implications of curating data (or failing to do so) when groups or family ties dissolve?

Our work also largely focuses on personal data that people explicitly create or acquire, for example photos, documents, messages, applications. These are types of data that people are familiar with and accustomed to interact with using established visual metaphors and paradigms (Chapter 7). But personal data now includes a growing number of types, with new devices generating more types. Future work can shift the focus from the types of data we explored to less visible data, such as audio recordings stored by voice assistants, data stored on social networks, and interactions recorded within *Internet of Things* devices.

An additional way for future research to leverage and extend our work is to expand our methodological approach. Our user modeling work is purely qualitative, but, as we mention in Chapter 6, future studies can use a quantitative approach to triangulate our results. Our design work in Chapter 7 and Chapter 8, instead, relies on prototypes and concepts, but future work could shift the focus to *research products* (i.e., "inquiry-driven," polished artifacts meant to be engaged with as they are) [216], to investigate long-term curation practices through field deployments. We can imagine ideas from some of the concepts in Chapter 7 or the prototype of Data Dashboard (Chapter 8) becoming part of a functional research product that can generate new insights about data curation practices over the long-term. Longitudinal studies could help better uncover the effects of engaging in personal data curation over time and unpack its temporal dimensions.

9.2.3 Implications for algorithm design

The design ideas we explore in the dissertation highlight the need for a parallel technical effort to make them viable. Data Dashboard in Chapter 8 shows how clustering personal data into automatic categories and types can be helpful, but a similar approach will only be as good as the algorithms behind it. Current management systems and platforms can recognize different types of data, but they largely rely on file extensions to do so. Instead, our work shows that often the content of data is the key focus of interest in curation. Improvements in automatic content classification and categorization algorithms can go a long way in supporting data curation.

However, given the sensitive nature of much of personal data, a key challenge is to make similar algorithms work entirely offline on local devices. Often online platforms can offer advanced functionalities for recognizing and sorting data because they rely on the cloud infrastructure behind them. But this might not be the best approach for supporting curation, as participants' reactions to Data Dashboard show (Chapter 8). Companies such as Google are exploring how to make machine learning available locally, entirely independent from cloud infrastructure, to preserve user privacy [162]. We believe that technical efforts for supporting personal data curation should focus on this same challenge.

9.2.4 Implications for policy and sustainability

Our work can also have implications for policy and sustainability.

Recent years have seen an increased focus on regulatory efforts for the storing and processing of personal data [1, 265]. However, our work suggests that any potential tools and policies in this domain should also address the confusing technical language of personal data that everyday users often struggle with (Chapter 8). Additional regulatory efforts should also shift focus from storing and retention practices to the algorithms behind advanced functionalities related to personal data. For example, they could focus on the algorithms for automatic recognition and categorization of personal data that we discuss in the previous section. Yes, we do think that similar algorithms can help, but there is a need for stronger regulation around who can use them and for what purposes.

Finally, an indirect but important implication of our work is the need to reflect on the possible environmental effects of storing and preserving large quantities of data. We are not Luddites. We do not advocate for a return to the days of paper-based personal information management. But storing digital data can require large quantities of energy [141, 199]. Although recent work shows that predictions about data centres' energy use have been largely exaggerated, there is still a need to closely monitor electricity consumption over the next years and proactively prepare for a growing demand in cloud data storage [185]. Making data storage a carbon-neutral practice is a first step, but often hard to achieve at scale [98, 272]. As the climate crisis exacerbates, there will be increasing pressure to ensure that data storing practices remain sustainable.

9.3 Closing remarks

Four years ago, the idea of automatically deleting data seemed extreme, a provocation. Now, it is almost an industry standard for products that collect sensitive data such as audio recordings and location history [40, 235].

Four years ago, the term "digital hoarding" was only briefly mentioned in a medical case study. Now, there is a growing body of research about keeping and discarding data practices, with perspectives from computer science, psychology, information science, and health sciences [171, 205, 206, 222, 253, 271].

In the immediate future, people's personal data will keep growing, with external actors wanting more and more of it, for uses that might not reflect what people want. For sure, this feels as the right time to take more control of personal data. But then, what to keep of it?

Bibliography

[1] Guide to the General Data Protection Regulation, May 2018. URL https://www.gov.uk/government/publications/guide-to-the-general-data-protection-regulation. → pages 1, 162, 173

[2] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015. → pages 133, 134

[3] A. Adya, W. J. Bolosky, M. Castro, G. Cermak, R. Chaiken, J. R. Douceur, J. Howell, J. R. Lorch, M. Theimer, and R. P. Wattenhofer. Farsite: Federated, Available, and Reliable Storage for an Incompletely Trusted Environment. *SIGOPS Oper. Syst. Rev.*, 36(SI):1–14, Dec. 2003. ISSN 0163-5980. doi:10.1145/844128.844130. → pages 164

[4] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. Tweets Are Forever: A Large-Scale Quantitative Analysis of Deleted Tweets. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, page 897–908, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450313315. doi:10.1145/2441776.2441878. → pages 17

[5] L. Alon and R. Nachmias. Gaps between Actual and Ideal Personal Information Management Behavior. *Computers in Human Behavior*, page 106292, 2020. → pages 16, 165

[6] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi. Differential Privacy: on the trade-off between Utility and Information Leakage. Mar. 2011. doi:10.1007/978-3-642-29420-4_3. URL https://arxiv.org/abs/1103.5188v3. → pages 155

[7] A. N. Antle. Child-Personas: Fact or Fiction? In *Proceedings of the 6th Conference on Designing Interactive Systems*, DIS '06, page 22–30, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933670. doi:10.1145/1142405.1142411. → pages 77

[8] Apple. How to use Stacks on your Mac, Sept. 2018. URL https://support.apple.com/en-us/HT209101. publisher: Apple. → pages 3

[9] C. Arthur. Naked celebrity hack: security experts focus on iCloud backup theory — Technology — The Guardian, Sep 2014. URL https://www.theguardian.com/technology/2014/sep/01/naked-celebrity-hack-icloud-backup-jennifer-lawrence. → pages 4

[10] R. Banks, D. Kirk, and A. Sellen. A design perspective on three technology heirlooms. *Human–Computer Interaction*, 27(1-2):63–91, 2012. → pages 15

[11] L. J. Bannon. Forgetting as a feature, not a bug: the duality of memory and implications for ubiquitous computing. *CoDesign*, 2(01):3–15, 2006. → pages 2

[12] N. Banovic, F. Chevalier, T. Grossman, and G. Fitzmaurice. Triggering Triggers and Burying Barriers to Customizing Software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 2717–2726, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi:10.1145/2207676.2208666. → pages 134, 135

[13] J. Bardram, J. Bunde Pedersen, and M. Soegaard. Support for Activity-based Computing in a Personal Computing Operating System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 211–220, New York, NY, USA, 2006. ACM. ISBN 1-59593-372-7. doi:10.1145/1124772.1124805. → pages 100, 101

[14] D. Barreau and B. A. Nardi. Finding and reminding: file organization from the desktop. *ACM SigChi Bulletin*, 27(3):39–43, 1995. → pages 10, 112

[15] S. Barth and M. D. De Jong. The privacy paradox–Investigating discrepancies between expressed privacy concerns and actual online

behavior–A systematic literature review. *Telematics and Informatics*, 34 (7):1038–1058, 2017. → pages 133, 134

[16] R. Bekkerman. Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora. 2004. → pages 160

[17] R. W. Belk. Possessions and the extended self. *Journal of consumer research*, 15(2):139–168, 1988. → pages 14, 91, 133

[18] R. W. Belk. Extended self in a digital world. *Journal of Consumer Research*, 40(3):477–500, 2013. → pages 14, 91

[19] V. Bellotti, N. Ducheneaut, M. Howard, and I. Smith. Taking Email to Task: The Design and Evaluation of a Task Management Centered Email Tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 345–352, New York, NY, USA, 2003. ACM. ISBN 1-58113-630-7. doi:10.1145/642611.642672. → pages 102, 132, 133

[20] D. Ben Menachem. Behavioral Archetypes, Feb. 2016. URL https://smashingideas.com/behavioral-archetypes/. → pages 77, 85

[21] D. Benyon. *Designing interactive systems: A comprehensive guide to HCI, UX and interaction design*. Pearson Edinburgh, 2014. → pages 143

[22] O. Bergman. The user-subjective approach to personal information management: from theory to practice. In *Human-computer interaction: the agency perspective*, pages 55–81. Springer, 2012. → pages 72, 101, 131

[23] O. Bergman. Variables for personal information management research. In *Aslib Proceedings: New Information Perspectives*, volume 65, pages 464–483. Emerald Group Publishing Limited, 2013. → pages 4, 70, 71, 72, 76, 77

[24] O. Bergman and S. Whittaker. *The Science of Managing Our Digital Stuff*. MIT Press, 2016. → pages 3, 14, 15, 16, 72, 73, 99

[25] O. Bergman and N. Yanai. Personal Information Retrieval: Smartphones vs. Computers, Emails vs. Files. *Personal Ubiquitous Comput.*, 22(4): 621–632, Aug. 2018. ISSN 1617-4909. doi:10.1007/s00779-017-1101-6. → pages 3, 72

[26] O. Bergman, R. Beyth Marom, and R. Nachmias. The user-subjective approach to personal information management systems. *Journal of the American Society for Information Science and Technology*, 54(9):872–878, 2003. doi:10.1002/asi.10283. → pages 2, 72, 102

[27] O. Bergman, R. Boardman, J. Gwizdka, and W. Jones. Personal Information Management. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pages 1598–1599, New York, NY, USA, 2004. ACM. ISBN 1-58113-703-6. doi:10.1145/985921.986164. → pages 9

[28] O. Bergman, R. Beyth Marom, R. Nachmias, N. Gradovitch, and S. Whittaker. Improved Search Engines and Navigation Preference in Personal Information Management. *ACM Trans. Inf. Syst.*, 26(4): 20:1–20:24, Oct. 2008. ISSN 1046-8188. doi:10.1145/1402256.1402259. → pages 3, 72, 73

[29] O. Bergman, S. Tucker, R. Beyth Marom, E. Cutrell, and S. Whittaker. It's Not That Important: Demoting Personal Information of Low Subjective Importance Using GrayArea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 269–278, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi:10.1145/1518701.1518745. → pages 101, 102, 127, 132

[30] O. Bergman, A. Komninos, D. Liarokapis, and J. Clarke. You Never Call: Demoting Unused Contacts on Mobile Phones Using DMTR. *Personal Ubiquitous Comput.*, 16(6):757–766, Aug. 2012. ISSN 1617-4909. doi:10.1007/s00779-011-0411-3. → pages 101, 102, 132

[31] O. Bergman, O. Elyada, N. Dvir, Y. Vaitzman, and A. B. Ami. Spotting the Latest Version of a File with Old'nGray. *Interacting with Computers*, 27 (6):630–639, 2015. → pages 101, 102, 127

[32] O. Bergman, S. Tucker, and S. Dahamshy. The effect of demoting near-duplicate pictures. *Proceedings of the Association for Information Science and Technology*, 55(1):755–756, 2018. → pages 132

[33] O. Bergman, T. Israeli, and S. Whittaker. Search is the future? The young search less for files. *Proceedings of the Association for Information Science and Technology*, 56(1):360–363, 2019. → pages 72

[34] O. Bergman, S. Whittaker, and Y. Frishman. Let's get personal: the little nudge that improves document retrieval in the Cloud. *Journal of Documentation*, 75(2):379–396, 2019. → pages 132

[35] J. D. Biersdorfer. New Ways to Delete Old Files - The New York Times, May 2018. URL https://www.nytimes.com/2018/05/17/technology/personaltech/new-ways-to-delete-old-files.html. → pages 66

[36] J. Blom. Personalization: A Taxonomy. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '00, page 313–314, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132484. doi:10.1145/633292.633483. → pages 134

[37] A. Blomquist and M. Arvola. Personas in Action: Ethnography in an Interaction Design Team. In *Proceedings of the Second Nordic Conference on Human-computer Interaction*, NordiCHI '02, pages 197–200, New York, NY, USA, 2002. ACM. ISBN 1-58113-616-1. doi:10.1145/572020.572044. → pages 77

[38] R. Boardman and M. A. Sasse. "Stuff Goes into the Computer and Doesn'T Come out": A Cross-tool Study of Personal Information Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 583–590, New York, NY, USA, 2004. ACM. ISBN 1-58113-702-8. doi:10.1145/985692.985766. → pages 2, 10, 15, 16, 72

[39] S. Bowen and D. Petrelli. Remembering today tomorrow: Exploring the human-centred design of digital mementos. *International Journal of Human-Computer Studies*, 69(5):324–337, 2011. → pages 15

[40] R. Brandom. Amazon pushes Alexa privacy with new delete options, Sept. 2019. URL https://www.theverge.com/2019/9/25/20883745/amazon-alexa-privacy-hub-security-voice-recordings-echo-devices. → pages 174

[41] V. Braun and V. Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, 2019. → pages 21, 25

[42] R. N. Brewer, M. R. Morris, and S. E. Lindley. How to Remember What to Remember: Exploring Possibilities for Digital Reminder Systems. *Proc.*

ACM Interact. Mob. Wearable Ubiquitous Technol., 1(3):38:1–38:20, Sept. 2017. ISSN 2474-9567. doi:10.1145/3130903. → pages 103

[43] M. Broekhuijsen, E. van den Hoven, and P. Markopoulos. From PhotoWork to PhotoUse: exploring personal digital photo activities. *Behaviour & Information Technology*, 36(7):754–767, 2017. → pages 14

[44] M. J. Broekhuijsen. *Curation-in-Action: design for photo curation to support shared remembering*. PhD thesis, 2018. → pages 14

[45] M. Buchenau and J. F. Suri. Experience Prototyping. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '00, pages 424–433, New York, NY, USA, 2000. ACM. ISBN 1-58113-219-0. doi:10.1145/347642.347802. → pages 20, 106

[46] A. Bunt, C. Conati, and J. McGrenere. Supporting Interface Customization Using a Mixed-initiative Approach. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 92–101, New York, NY, USA, 2007. ACM. ISBN 1-59593-481-2. doi:10.1145/1216295.1216317. → pages 109, 134, 135

[47] C. Cadwalladr and E. Graham Harrison. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach — The Guardian, March 2018. URL https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. → pages 127

[48] A. Calero Valdez and M. Ziefle. The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies*, 121:108–121, Jan. 2019. ISSN 10715819. doi:10.1016/j.ijhcs.2018.04.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1071581918301642. → pages 155

[49] R. Capra, J. Khanova, and S. Ramdeen. Work and personal e-mail use by university employees: PIM practices across domain boundaries. *Journal of the American Society for Information Science and Technology*, 64(5): 1029–1044, 2013. → pages 133

[50] N. Carr. I am a data factory (and so are you), 2018. URL http://www.roughtype.com/?p=8394. → pages 2

[51] M. E. Cecchinato, A. L. Cox, and J. Bird. Working 9-5? Professional Differences in Email and Boundary Management Practices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3989–3998, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi:10.1145/2702123.2702537. → pages 133

[52] S. Cerf, V. Primault, A. Boutet, S. B. Mokhtar, R. Birke, S. Bouchenak, L. Y. Chen, N. Marchand, and B. Robu. PULP: Achieving Privacy and Utility Trade-Off in User Mobility Data. In *2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS)*, pages 164–173, Sept. 2017. doi:10.1109/SRDS.2017.25. → pages 155

[53] C. N. Chapman and R. P. Milham. The personas' new clothes: methodological and practical arguments against a popular method. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 634–636. SAGE Publications Sage CA: Los Angeles, CA, 2006. → pages 93

[54] K. Charmaz. The search for Meanings–Grounded Theory. In. Smith JA, Harre R., & Van Langenhove L.(Eds.), Rethinking Methods in Psychology (pp. 27–49), 1996. → pages 20

[55] K. Charmaz. Constructionism and the grounded theory method. *Handbook of constructionist research*, 1:397–412, 2008. → pages 20

[56] K. Charmaz. *Constructing grounded theory*. Sage, 2014. → pages 19, 20, 25, 34, 55, 56, 73, 76

[57] A. Y. S. Chen, W. Odom, C. Zhong, H. Lin, and T. Amram. Chronoscope: Designing Temporally Diverse Interactions with Personal Digital Photo Collections. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, page 799–812, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450358507. doi:10.1145/3322276.3322301. → pages 95

[58] E. Cheon and N. M. Su. "Staged for Living": Negotiating Objects and Their Values over a Porous Boundary. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), Nov. 2018. doi:10.1145/3274305. → pages 55, 133, 165

[59] E. Cheon and N. M. Su. The Value of Empty Space for Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 49:1–49:13, New York, NY, USA, 2018. ACM.

ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3173623. → pages 15, 55, 165

[60] H. Cherrier and J. B. Murray. Reflexive dispossession and the self: constructing a processual theory of identity. *Consumption Markets & Culture*, 10(1):1–29, 2007. → pages 15

[61] V. Clarke and V. Braun. Thematic analysis. In *Encyclopedia of critical psychology*, pages 1947–1952. Springer, 2014. → pages 21, 32, 76, 116, 148

[62] V. Clarke and V. Braun. Questions about thematic analysis - The University of Auckland, n.d. URL https://www.psych.auckland.ac.nz/en/about/our-research/research-groups/thematic-analysis/frequently-asked-questions-8.html. → pages 32, 34

[63] M. Crotty. *The foundations of social research: Meaning and perspective in the research process*. Sage, 1998. → pages 22

[64] G. Cselle, K. Albrecht, and R. Wattenhofer. BuzzTrack: topic detection and tracking in email. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 190–197. ACM, 2007. → pages 160

[65] M. Csikszentmihalyi and E. Halton. *The meaning of things: Domestic symbols and the self*. Cambridge University Press, 1981. → pages 14, 91

[66] A. L. Cushing. Self extension and the desire to preserve digital possessions. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–3, 2011. → pages 14, 29, 47, 91

[67] A. L. Cushing. *Possessions and self extension in digital environments: Implications for maintaining personal information*. PhD thesis, The University of North Carolina at Chapel Hill, 2012. → pages

[68] A. L. Cushing. "It's stuff that speaks to me": Exploring the characteristics of digital possessions. *Journal of the Association for Information Science and Technology*, 64(8):1723–1734, 2013. → pages 14, 15, 29, 47, 91

[69] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin. Fast, Flexible Filtering with Phlat. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 261–270, New York, NY, USA, 2006. ACM. ISBN 1-59593-372-7. doi:10.1145/1124772.1124812. → pages 101, 102, 132, 140

[70] K. Cutting and E. Hedenborg. Can Personas Speak? Biopolitics in Design Processes. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, DIS '19 Companion, page 153–157, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362702. doi:10.1145/3301019.3323911. → pages 93

[71] F. Dabek, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica. Wide-Area Cooperative Storage with CFS. *SIGOPS Oper. Syst. Rev.*, 35(5):202–215, Oct. 2001. ISSN 0163-5980. doi:10.1145/502059.502054. → pages 164

[72] S. Davidoff, M. K. Lee, A. K. Dey, and J. Zimmerman. Rapidly exploring application design through speed dating. In *International Conference on Ubiquitous Computing*, pages 429–446. Springer, 2007. → pages 20, 106

[73] J. D. Dinneen and C.-A. Julien. What's in people's digital file collections? *Proceedings of the Association for Information Science and Technology*, 56 (1):68–77, 2019. → pages 2, 144

[74] J. D. Dinneen and C.-A. Julien. The ubiquitous digital file: A review of file management research. *Journal of the Association for Information Science and Technology*, 2019. → pages 10

[75] J. D. Dinneen, F. Odoni, I. Frissen, and C.-A. Julien. Cardinal: Novel software for studying file management behavior. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology*, page 62. American Society for Information Science, 2016. → pages 140

[76] J. D. Dinneen, C.-A. Julien, and I. Frissen. The Scale and Structure of Personal File Collections. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 327:1–327:12, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2. doi:10.1145/3290605.3300557. → pages 2, 144

[77] D. Dion, O. Sabri, and V. Guillard. Home sweet messy home: Managing symbolic pollution. *Journal of Consumer Research*, 41(3):565–589, 2014. → pages 133

[78] P. Dourish. Implications for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 541–550, New York, NY, USA, 2006. Association for Computing

Machinery. ISBN 1595933727. doi:10.1145/1124772.1124855. → pages 169

[79] P. Dourish, W. K. Edwards, A. LaMarca, and M. Salisbury. Presto: An Experimental Architecture for Fluid Interactive Document Spaces. *ACM Trans. Comput.-Hum. Interact.*, 6(2):133–161, June 1999. ISSN 1073-0516. doi:10.1145/319091.319099. → pages 101, 102, 132

[80] Dropbox. How to create and share Google Docs, Sheets, and Slides in Dropbox, NA. URL https://help.dropbox.com/installs-integrations/third-party/create-google-docs. → pages 159

[81] Dropbox. How to use selective sync, NA. URL https://help.dropbox.com/installs-integrations/sync-uploads/selective-sync-overview. → pages 140

[82] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff I'Ve Seen: A System for Personal Information Retrieval and Re-Use. *SIGIR Forum*, 49(2):28–35, Jan. 2016. ISSN 0163-5840. doi:10.1145/2888422.2888425. → pages 101, 102, 132, 140

[83] J. L. Dupree, R. Devries, D. M. Berry, and E. Lank. Privacy Personas: Clustering Users via Attitudes and Behaviors toward Security Practices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5228–5239, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858214. → pages 77

[84] A. K. Dutta and R. Hasan. How much does storage really cost? Towards a full cost accounting model for data storage. In *International Conference on Grid Economics and Business Models*, pages 29–43. Springer, 2013. → pages 2

[85] C. Elsden, D. S. Kirk, and A. C. Durrant. A quantified past: Toward design for remembering with personal informatics. *Human–Computer Interaction*, 31(6):518–557, 2016. → pages 2, 11, 59

[86] C. Elsden, M. Selby, A. Durrant, and D. Kirk. Fitter, happier, more productive: what to ask of a data-driven life. *Interactions*, 23(5):45–45, 2016. → pages 162

[87] D. A. Epstein, A. Ping, J. Fogarty, and S. A. Munson. A Lived Informatics Model of Personal Informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 731–742, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3574-4. doi:10.1145/2750858.2804250. → pages 11, 59

[88] S. Faily and I. Flechais. Persona Cases: A Technique for Grounding Personas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 2267–2270, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi:10.1145/1978942.1979274. → pages 93

[89] M. Feinberg, G. Geisler, E. Whitworth, and E. Clark. Understanding Personal Digital Collections: An Interdisciplinary Exploration. In *Proceedings of the Designing Interactive Systems Conference*, DIS '12, page 200–209, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312103. doi:10.1145/2317956.2317988. → pages 14

[90] S. Fertig, E. Freeman, and D. Gelernter. Lifestreams: An Alternative to the Desktop Metaphor. In *Conference Companion on Human Factors in Computing Systems*, CHI '96, pages 410–411, New York, NY, USA, 1996. ACM. ISBN 0-89791-832-0. doi:10.1145/257089.257404. → pages 101, 102, 132

[91] L. Findlater and J. McGrenere. A Comparison of Static, Adaptive, and Adaptable Menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 89–96, New York, NY, USA, 2004. ACM. ISBN 1-58113-702-8. doi:10.1145/985692.985704. → pages 102

[92] L. Findlater and J. McGrenere. Beyond performance: Feature awareness in personalized interfaces. *International Journal of Human-Computer Studies*, 68(3):121–137, 2010. → pages 135

[93] S. Fitchett, A. Cockburn, and C. Gutwin. Finder Highlights: Field Evaluation and Design of an Augmented File Browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3685–3694, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi:10.1145/2556288.2557014. → pages 101, 102

[94] A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al. Above the clouds: A berkeley view of cloud computing. *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS*, 28(13):2009, 2009. → pages 163

[95] R. O. Frost, G. Steketee, and J. Grisham. Measurement of compulsive hoarding: saving inventory-revised. *Behaviour Research and Therapy*, 42 (10):1163–1182, 2004. → pages 30

[96] W. Gaver. What Should We Expect from Research through Design? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 937–946, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi:10.1145/2207676.2208538. → pages 20

[97] R. Geambasu, T. Kohno, A. A. Levy, and H. M. Levy. Vanish: Increasing Data Privacy with Self-Destructing Data. In *USENIX Security Symposium*, volume 316, 2009. → pages 67, 101, 102

[98] M. Geuss. Five graphics from Google show how carbon-intensive its data centers really are, Oct. 2018. URL https://arstechnica.com/information-technology/2018/10/googles-data-center-carbon-heat-maps-show-the-challenges-of-going-carbon-free/. → pages 173

[99] E. Giaccardi, E. Karana, H. Robbins, and P. D'Olivo. Growing Traces on Objects of Daily Use: A Product Design Perspective for HCI. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, DIS '14, pages 473–482, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2902-6. doi:10.1145/2598510.2602964. → pages 66, 108

[100] S. Gibbons. Nielsen Norman Group: UX Research, Training, and Consulting, Jan. 2018. URL https://www.nngroup.com/articles/empathy-mapping/. → pages 77

[101] G. Gibbs. A Discussion with Prof Kathy Charmaz on Grounded Theory, Feb. 2015. URL https://www.youtube.com/watch?v=D5AHmHQS6WQ. → pages 25

[102] A. Girgensohn, S. A. Bly, F. Shipman, J. S. Boreczky, and L. Wilcox. Home Video Editing Made Easy-Balancing Automation and User Control. In *INTERACT*, volume 1, pages 464–471, 2001. → pages 102

[103] B. G. Glaser and A. L. Strauss. Grounded theory: Strategies for qualitative research. *Chicago, lL: Aldine Publishing Company*, 1967. → pages 20

[104] C. Golsteijn, E. van den Hoven, D. Frohlich, and A. Sellen. Towards a More Cherishable Digital Object. In *Proceedings of the Designing Interactive Systems Conference*, DIS '12, pages 655–664, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1210-3. doi:10.1145/2317956.2318054. → pages 15, 36

[105] C. J. Gormley and S. J. Gormley. Data hoarding and Information Clutter: The Impact on Cost, Life Span of Data, Effectivness, Sharing, Productivity, and Knowledge Management Culture. *Issues in Information Systems*, 13: 90–95, 2012. → pages 30

[106] S. Greenberg and I. H. Witten. Adaptive personalized interfaces—A question of viability. *Behaviour & Information Technology*, 4(1):31–45, 1985. → pages 135

[107] G. Greenwald and E. MacAskill. NSA Prism program taps in to user data of Apple, Google and others — The Guardian, June 2013. URL https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data. → pages 127

[108] J. Gruning. Displaying Invisible Objects: Why People Rarely Re-read E-books. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 139:1–139:12, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3173713. → pages 133

[109] J. Gruning and S. Lindley. Things We Own Together: Sharing Possessions at Home. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1176–1186, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi:10.1145/2858036.2858154. → pages 64

[110] E. G. Guba. Criteria for assessing the trustworthiness of naturalistic inquiries. *Ectj*, 29(2):75, 1981. → pages 22, 23, 96

[111] R. Gulotta, W. Odom, J. Forlizzi, and H. Faste. Digital Artifacts As Legacy: Exploring the Lifespan and Value of Digital Data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1813–1822, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1899-0. doi:10.1145/2470654.2466240. → pages 2, 10, 14, 15, 65, 99, 124, 133, 143, 144

[112] R. Gulotta, A. Sciuto, A. Kelliher, and J. Forlizzi. Curatorial Agents: How Systems Shape Our Understanding of Personal and Familial Digital Information. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3453–3462, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi:10.1145/2702123.2702297. → pages 14, 102, 104, 133, 143

[113] A. Gutmann and M. Warner. Fight to Be Forgotten: Exploring the Efficacy of Data Erasure in Popular Operating Systems. In *Annual Privacy Forum*, pages 45–58. Springer, 2019. → pages 162

[114] J. Gwizdka. Timely Reminders: A Case Study of Temporal Guidance in PIM and Email Tools Usage. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '00, pages 163–164, New York, NY, USA, 2000. ACM. ISBN 1-58113-248-4. doi:10.1145/633292.633383. → pages 103

[115] J. Gwizdka. Supporting Prospective Information in Email. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, pages 135–136, New York, NY, USA, 2001. ACM. ISBN 1-58113-340-5. doi:10.1145/634067.634150. → pages

[116] J. Gwizdka. Reinventing the Inbox: Supporting the Management of Pending Tasks in Email. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, pages 550–551, New York, NY, USA, 2002. ACM. ISBN 1-58113-454-1. doi:10.1145/506443.506476. → pages 103

[117] J. Gwizdka. Email Task Management Styles: The Cleaners and the Keepers. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pages 1235–1238, New York, NY, USA, 2004. ACM. ISBN 1-58113-703-6. doi:10.1145/985921.986032. → pages 3, 10, 29, 48, 72

[118] J. Gwizdka and M. Chignell. Individual Differences. In W. Jones and J. Teevan, editors, *Personal Information Management*, chapter 12, pages

206–220. University of Washington Press, Seattle and London, 2007. → pages 4, 70, 72

[119] M. Haraty and J. McGrenere. Designing for Advanced Personalization in Personal Task Management. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, page 239–250, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340311. doi:10.1145/2901790.2901805. → pages 132, 134, 135, 142

[120] M. Haraty, J. McGrenere, and A. Bunt. Online Customization Sharing Ecosystems: Components, Roles, and Motivations. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 2359–2371, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343350. doi:10.1145/2998181.2998289. → pages 161

[121] M. Haraty, Z. Wang, H. Wang, S. Iqbal, and J. Teevan. Design and in-situ evaluation of a mixed-initiative approach to information organization. *Journal of the Association for Information Science and Technology*, 68(9): 2211–2224, 2017. → pages 134, 135

[122] R. Harper and W. Odom. Trusting oneself: an anthropology of digital things and personal competence. *Trust, Computing, and Society*, page 272, 2014. → pages 15

[123] R. Harper, S. Lindley, E. Thereska, R. Banks, P. Gosset, G. Smyth, W. Odom, and E. Whitworth. What is a File? In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1125–1136, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1331-5. doi:10.1145/2441776.2441903. → pages 10, 126, 162

[124] S. Harrison, D. Tatar, and P. Sengers. The three paradigms of HCI. In *Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems San Jose, California, USA*, pages 1–18, 2007. → pages 34

[125] B. Hecht, L. Wilcox, J. Bigham, J. Schöning, E. Hoque, J. Ernst, Y. Bisk, L. De Russis, L. Yarosh, B. Anjum, D. Contractor, and C. Wu. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process, March 2018. URL https://acm-fca.org/2018/03/29/negativeimpacts/. → pages 128

[126] S. Henderson. *How do people manage their documents?: an empirical investigation into personal document management practices among knowledge workers*. PhD thesis, ResearchSpace@ Auckland, 2009. → pages 30, 48

[127] S. Henderson. Personal Document Management Strategies. In *Proceedings of the 10th International Conference NZ Chapter of the ACM's Special Interest Group on Human-Computer Interaction*, CHINZ '09, pages 69–76, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-574-1. doi:10.1145/1577782.1577795. → pages 3, 30, 48

[128] S. Henderson and A. Srinivasan. Filing, piling & structuring: strategies for personal document management. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011. → pages 72

[129] L. A. Henkel. Point-and-shoot memories: The influence of taking photos on memory for a museum tour. *Psychological science*, 25(2):396–402, 2014. → pages 49

[130] D. Herron, W. Moncur, and E. van den Hoven. Digital Decoupling and Disentangling: Towards Design for Romantic Break Up. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, page 1175–1185, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349222. doi:10.1145/3064663.3064765. → pages 15

[131] C. G. Hill, M. Haag, A. Oleson, C. Mendez, N. Marsden, A. Sarma, and M. Burnett. Gender-Inclusiveness Personas vs. Stereotyping: Can We Have It Both Ways? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 6658–6671, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi:10.1145/3025453.3025609. → pages 93

[132] W. C. Hill, J. D. Hollan, D. Wroblewski, and T. McCandless. Edit wear and read wear. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 3–9. ACM, 1992. → pages 66

[133] V. Hollis, A. Konrad, A. Springer, M. Antoun, C. Antoun, R. Martin, and S. Whittaker. What does all this data mean for my future mood? Actionable Analytics and Targeted Reflection for Emotional Well-Being. *Human–Computer Interaction*, 32(5-6):208–267, 2017. → pages 2

[134] E. Horvitz. Principles of Mixed-initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 159–166, New York, NY, USA, 1999. ACM. ISBN 0-201-48559-1. doi:10.1145/302979.303030. → pages 109, 134

[135] A. Hsu. How Forgetting Might Make Us Smarter - NPR, June 2017. URL http://www.npr.org/sections/health-shots/2017/06/23/534001592/could-the-best-memory-system-be-one-that-forgets. → pages 49

[136] J. Huh, M. Pollack, H. Katebi, K. Sakallah, and N. Kirsch. Incorporating User Control in Automated Interactive Scheduling Systems. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, pages 306–309, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0103-9. doi:10.1145/1858171.1858226. → pages 102

[137] A. Hurst, J. Mankoff, A. K. Dey, and S. E. Hudson. Dirty Desktops: Using a Patina of Magnetic Mouse Dust to Make Common Interactor Targets Easier to Select. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, UIST '07, pages 183–186, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-679-0. doi:10.1145/1294211.1294242. → pages 108

[138] I. Ion, N. Sachdeva, P. Kumaraguru, and S. Čapkun. Home is Safer Than the Cloud!: Privacy Concerns for Consumer Cloud Storage. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 13:1–13:20, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0911-0. doi:10.1145/2078827.2078845. → pages 28

[139] R. Jenkins. Categorization: Identity, social process and epistemology. *Current sociology*, 48(3):7–25, 2000. → pages 91

[140] J. Jones and M. S. Ackerman. Curating an Infinite Basement: Understanding How People Manage Collections of Sentimental Artifacts. In *Proceedings of the 19th International Conference on Supporting Group Work*, GROUP '16, pages 87–97, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4276-6. doi:10.1145/2957276.2957316. → pages 14, 15, 57, 64, 91, 102, 103, 104, 133

[141] N. Jones. How to stop data centres from gobbling up the world's electricity. *Nature*, 561:163–166, Sept. 2018. doi:10.1038/d41586-018-06610-y. URL

https://www.nature.com/articles/d41586-018-06610-y. → pages 173

[142] W. Jones. Personal information management. *Annual review of information science and technology*, 41(1):453–504, 2007. → pages 9, 102

[143] W. Jones, P. Klasnja, A. Civan, and M. L. Adcock. The Personal Project Planner: Planning to Organize Personal Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 681–684, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1. doi:10.1145/1357054.1357162. → pages 101, 102, 132, 140, 153

[144] W. Jones, V. Bellotti, R. Capra, J. D. Dinneen, G. Mark, C. Marshall, K. Moffatt, J. Teevan, and M. Van Kleek. For Richer, for Poorer, in Sickness or in Health...: The Long-Term Management of Personal Information. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 3508–3515, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4082-3. doi:10.1145/2851581.2856481. → pages 2, 69, 102

[145] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013. → pages 3, 15, 126

[146] M. Kallahalla, E. Riedel, R. Swaminathan, Q. Wang, and K. Fu. Plutus: Scalable Secure File Sharing on Untrusted Storage. In *Proceedings of the 2nd USENIX Conference on File and Storage Technologies*, FAST '03, page 29–42, USA, 2003. USENIX Association. → pages 164

[147] V. Kalnikaitė and S. Whittaker. A saunter down memory lane: Digital reflection on personal mementos. *International Journal of Human-Computer Studies*, 69(5):298–310, 2011. → pages 15

[148] V. Kalnikaite, A. Sellen, S. Whittaker, and D. Kirk. Now Let Me See Where i Was: Understanding How Lifelogs Mediate Memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 2045–2054, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589299. doi:10.1145/1753326.1753638. → pages 11

[149] J. J. Kaye, J. Vertesi, S. Avery, A. Dafoe, S. David, L. Onaga, I. Rosero, and T. Pinch. To Have and to Hold: Exploring the Personal Archive. In

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06, pages 275–284, New York, NY, USA, 2006. ACM. ISBN 1-59593-372-7. doi:10.1145/1124772.1124814. → pages 14, 16, 29, 30, 47, 49, 91, 133

[150] M. T. Khan, M. Hyun, C. Kanich, and B. Ur. Forgotten But Not Gone: Identifying the Need for Longitudinal Data Management in Cloud Storage. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 543:1–543:12, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3174117. → pages 4, 16, 17, 65, 70, 99, 131

[151] S. Kim. *Personal digital archives: preservation of documents, preservation of self*. PhD thesis, 2013. → pages 16, 62, 64, 65, 95, 103

[152] D. Kirk, A. Sellen, C. Rother, and K. Wood. Understanding Photowork. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 761–770, New York, NY, USA, 2006. ACM. ISBN 1-59593-372-7. doi:10.1145/1124772.1124885. → pages 14, 62

[153] D. S. Kirk and A. Sellen. On Human Remains: Values and Practice in the Home Archiving of Cherished Objects. *ACM Trans. Comput.-Hum. Interact.*, 17(3), July 2010. ISSN 1073-0516. doi:10.1145/1806923.1806924. → pages 14, 15, 133

[154] D. S. Kirk, S. Izadi, A. Sellen, S. Taylor, R. Banks, and O. Hilliges. Opening up the Family Archive. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, page 261–270, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587950. doi:10.1145/1718918.1718968. → pages 133

[155] A. Klein. Hard Drive Cost Per Gigabyte, July 2017. URL https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/. → pages 2

[156] M. Kljun, J. Mariani, and A. Dix. Toward understanding short-term personal information preservation: A study of backup strategies of end users. *Journal of the Association for Information Science and Technology*, 67(12):2947–2963, 2016. → pages 29, 30

[157] A. Konrad, E. Isaacs, and S. Whittaker. Technology-mediated memory: Is technology altering our memories and interfering with well-being? *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4):23, 2016. → pages 2

[158] M. W. Lansdale. The psychology of personal information management. *Applied ergonomics*, 19(1):55–66, 1988. → pages 9, 10, 72

[159] J. L. Lastovicka and K. V. Fernandez. Three paths to disposition: The movement of meaningful possessions to strangers. *Journal of Consumer Research*, 31(4):813–823, 2005. → pages 15

[160] M.-H. Lee, S. Cha, and T.-J. Nam. Patina Engraver: Visualizing Activity Logs As Patina in Fashionable Trackers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1173–1182, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi:10.1145/2702123.2702213. → pages 108

[161] M.-H. Lee, O. Son, and T.-J. Nam. Patina-inspired Personalization: Personalizing Products with Traces of Daily Use. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, pages 251–263, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4031-1. doi:10.1145/2901790.2901812. → pages 108

[162] T. B. Lee. Why Google believes machine learning is its future, May 2019. URL https://arstechnica.com/gadgets/2019/05/googles-machine-learning-strategy-hardware-software-and-lots-of-data/. → pages 173

[163] I. Li, A. Dey, and J. Forlizzi. A Stage-based Model of Personal Informatics Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 557–566, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9. doi:10.1145/1753326.1753409. → pages 11

[164] J. Li, J. H. Lim, and Q. Tian. Automatic summarization for personal digital photos. In *Proc. of ICICS-PCM*, volume 3. Citeseer, 2003. → pages 102

[165] Y. S. Lincoln and E. G. Guba. But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New directions for program evaluation*, 1986(30):73–84, 1986. → pages 23, 96

[166] S. E. Lindley, C. C. Marshall, R. Banks, A. Sellen, and T. Regan. Rethinking the Web As a Personal Archive. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 749–760, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2035-1. doi:10.1145/2488388.2488454. → pages 3, 16, 133, 138, 139, 154, 158

[167] S. E. Lindley, G. Smyth, R. Corish, A. Loukianov, M. Golembewski, E. A. Luger, and A. Sellen. Exploring New Metaphors for a Networked World Through the File Biography. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 118:1–118:12, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3173692. → pages 66, 101, 126

[168] W. Liu, O. Rioul, J. McGrenere, W. E. Mackay, and M. Beaudouin Lafon. BIGFile: Bayesian Information Gain for Fast File Retrieval. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 385:1–385:13, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3173959. → pages 101, 102

[169] C. Lo, J. Cheng, and J. Leskovec. Understanding Online Collection Growth Over Time: A Case Study of Pinterest. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 545–554, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4914-7. doi:10.1145/3041021.3054189. → pages 107

[170] V. Luckerson. Why Google Is Suddenly Obsessed With Your Photos - The Ringer, May 2017. URL https://www.theringer.com/2017/5/25/16043842/google-photos-data-collection-e8578b3256e0. → pages 50

[171] A. M. Luxon, C. E. Hamilton, S. Bates, and G. S. Chasson. Pinning our possessions: Associations between digital hoarding and symptoms of hoarding disorder. *Journal of Obsessive-Compulsive and Related Disorders*, 21:60–68, 2019. → pages 3, 174

[172] S. Machvovech. Amazon confirms that Echo device secretly shared user's private audio — Ars Technica, May 2018. URL

https://arstechnica.com/gadgets/2018/05/amazon-confirms-that-echo-device-secretly-shared-users-private-audio/. → pages 4

[173] W. E. Mackay. More Than Just a Communication System: Diversity in the Use of Electronic Mail. In *Proceedings of the 1988 ACM Conference on Computer-supported Cooperative Work*, CSCW '88, pages 344–353, New York, NY, USA, 1988. ACM. ISBN 0-89791-282-9. doi:10.1145/62266.62293. → pages 72

[174] W. E. Mackay. Triggers and Barriers to Customizing Software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, page 153–160, New York, NY, USA, 1991. Association for Computing Machinery. ISBN 0897913833. doi:10.1145/108844.108867. → pages 135

[175] W. E. Mackay and A.-L. Fayard. HCI, Natural Science and Design: A Framework for Triangulation across Disciplines. In *Proceedings of the 2nd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '97, page 223–234, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918630. doi:10.1145/263552.263612. → pages 21

[176] M. Madden and L. Rainie. Americans' Attitudes About Privacy, Security and Surveillance — Pew Research Center, May 2015. URL http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/. → pages 4, 127

[177] T. W. Malone. How Do People Organize Their Desks?: Implications for the Design of Office Information Systems. *ACM Trans. Inf. Syst.*, 1(1): 99–112, Jan. 1983. ISSN 1046-8188. doi:10.1145/357423.357430. → pages 9, 10, 29, 48, 72

[178] C. Mancini, Y. Rogers, A. K. Bandara, T. Coe, L. Jedrzejczyk, A. N. Joinson, B. A. Price, K. Thomas, and B. Nuseibeh. Contravision: Exploring Users' Reactions to Futuristic Technology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 153–162, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9. doi:10.1145/1753326.1753350. → pages 106

[179] N. Marsden and M. Haag. Stereotypes and Politics: Reflections on Personas. In *Proceedings of the 2016 CHI Conference on Human Factors*

in Computing Systems, CHI '16, page 4017–4031, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858151. → pages 92

[180] N. Marsden and M. Pröbster. Personas and Identity: Looking at Multiple Identities to Inform the Construction of Personas. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi:10.1145/3290605.3300565. → pages 92

[181] C. Marshall and G. B. Rossman. *Designing qualitative research*. Sage publications, 2014. → pages 32

[182] C. C. Marshall. Rethinking personal digital archiving, Part 1: Four challenges from the field. *D-Lib Magazine*, 14(3/4):2, 2008. → pages 3, 28, 29, 30, 99

[183] C. C. Marshall. Rethinking personal digital archiving, part 2: implications for services, applications, and institutions. *D-Lib Magazine*, 14(3):3, 2008. → pages 2, 3, 140

[184] C. C. Marshall, S. Bly, and F. Brun Cottan. The long term fate of our digital belongings: Toward a service model for personal archives. In *Archiving Conference*, volume 2006, pages 25–30. Society for Imaging Science and Technology, 2006. → pages 16, 28, 29, 30

[185] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey. Recalibrating global data center energy-use estimates. *Science*, 367(6481):984–986, 2020. ISSN 0036-8075. doi:10.1126/science.aba3758. URL https://science.sciencemag.org/content/367/6481/984. → pages 173

[186] C. Massey, T. Lennig, and S. Whittaker. Cloudy Forecast: An Exploration of the Factors Underlying Shared Repository Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2461–2470, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450324731. doi:10.1145/2556288.2557042. → pages 12

[187] C. Massey, S. TenBrook, C. Tatum, and S. Whittaker. PIM and Personality: What Do Our Personal File Systems Say About Us? In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3695–3704, New York, NY, USA, 2014. ACM. ISBN

978-1-4503-2473-1. doi:10.1145/2556288.2557023. → pages 72, 129

[188] J. Matejka, T. Grossman, and G. Fitzmaurice. Patina: Dynamic Heatmaps for Visualizing Application Usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3227–3236, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1899-0. doi:10.1145/2470654.2466442. → pages 108

[189] T. Matthews, T. Judge, and S. Whittaker. How Do Designers and User Experience Professionals Actually Perceive and Use Personas? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1219–1228, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi:10.1145/2207676.2208573. → pages 93

[190] N. Maudet. Dead Angles of Personalization: Integrating Curation Algorithms in the Fabric of Design. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, page 1439–1448, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450358507. doi:10.1145/3322276.3322322. → pages 134

[191] M. J. May, E. Laron, K. Zoabi, and H. Gerhardt. On the Lifecycle of the File. *ACM Trans. Storage*, 15(1), Feb. 2019. ISSN 1553-3077. doi:10.1145/3295463. → pages 10

[192] M. L. Mazurek, E. Thereska, D. Gunawardena, R. Harper, and J. Scott. ZZFS: A hybrid device and cloud file system for spontaneous users. 2012. → pages 67

[193] J. McGrenere, R. M. Baecker, and K. S. Booth. An Evaluation of a Multiple Interface Design Solution for Bloated Software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, page 164–170, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134533. doi:10.1145/503376.503406. → pages 134, 135

[194] G. H. Mead. *Mind, self and society*, volume 111. Chicago University of Chicago Press., 1934. → pages 91

[195] D. Merritt, J. Jones, M. S. Ackerman, and W. S. Lasecki. Kurator: Using The Crowd to Help Families With Personal Curation Tasks. In *Proceedings of the 2017 ACM Conference on Computer Supported*

Cooperative Work and Social Computing, CSCW '17, pages 1835–1849, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4335-0. doi:10.1145/2998181.2998358. → pages 102

[196] D. Miller. *The comfort of things*. Polity, 2008. → pages 15

[197] K. Moon and D. Blackman. A guide to understanding social science research for natural scientists. *Conservation Biology*, 28(5):1167–1177, 2014. → pages 19, 22, 34

[198] R. Mortier, J. Zhao, J. Crowcroft, L. Wang, Q. Li, H. Haddadi, Y. Amar, A. Crabtree, J. Colley, T. Lodge, T. Brown, D. McAuley, and C. Greenhalgh. Personal Data Management with the Databox: What's Inside the Box? In *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking*, CAN '16, pages 49–54, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4673-3. doi:10.1145/3010079.3010082. → pages 127

[199] S. Moss. Pointless emails: they're not just irritating – they have a massive carbon footprint. *The Guardian*, Nov. 2019. ISSN 0261-3077. URL https://www.theguardian.com/technology/shortcuts/2019/nov/26/pointless-emails-theyre-not-just-irritating-they-have-a-massive-carbon-footprint. → pages 173

[200] Mozilla. *Privacy Not Included: A Buyer's Guide for Connected Products, n.d. URL https://foundation.mozilla.org/en/privacynotincluded/. → pages 94

[201] M. J. Muller and S. Kogan. Grounded theory method in HCI and CSCW. *Cambridge: IBM Center for Social Software*, pages 1–46, 2010. → pages 19

[202] A. Murillo, A. Kramm, S. Schnorf, and A. De Luca. " If I press delete, it's gone"-User Understanding of Online Data Deletion and Expiration. In *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*, pages 329–339, 2018. → pages 17, 112

[203] J. Naughton. 'The goal is to automate us': welcome to the age of surveillance capitalism. *The Guardian*, Jan. 2019. ISSN 0029-7712. URL https://www.theguardian.com/technology/2019/jan/20/

shoshana-zuboff-age-of-surveillance-capitalism-google-facebook. → pages 162

[204] T. Neate, A. Bourazeri, A. Roper, S. Stumpf, and S. Wilson. Co-Created Personas: Engaging and Empowering Users with Diverse Needs Within the Design Process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi:10.1145/3290605.3300880. → pages 92

[205] N. Neave, P. Briggs, K. McKellar, and E. Sillence. Digital hoarding behaviours: Measurement and evaluation. *Computers in Human Behavior*, 96:72–77, 2019. → pages 3, 30, 174

[206] N. Neave, K. McKellar, E. Sillence, and P. Briggs. Digital hoarding behaviours: Implications for cybersecurity. In *Cyber Influence and Cognitive Threats*, pages 77–95. Elsevier, 2020. → pages 30, 174

[207] M. Nouwens, C. F. Griggio, and W. E. Mackay. "WhatsApp is for Family; Messenger is for Friends": Communication Places in App Ecosystems. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 727–735, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9. doi:10.1145/3025453.3025484. → pages 47, 133

[208] P. Obrador, R. de Oliveira, and N. Oliver. Supporting Personal Photo Storytelling for Social Albums. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 561–570, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi:10.1145/1873951.1874025. → pages 102, 132, 135

[209] W. Odom and T. Duel. On the Design of OLO Radio: Investigating Metadata As a Design Material. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 104:1–104:9, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3173678. → pages 95, 103

[210] W. Odom, J. Pierce, E. Stolterman, and E. Blevis. Understanding Why We Preserve Some Things and Discard Others in the Context of Interaction Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1053–1062, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi:10.1145/1518701.1518862. → pages 15, 95

[211] W. Odom, J. Zimmerman, and J. Forlizzi. Teenagers and Their Virtual Possessions: Design Opportunities and Issues. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1491–1500, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi:10.1145/1978942.1979161. → pages 10, 14, 16, 29, 47, 66, 108, 133

[212] W. Odom, R. Banks, D. Kirk, R. Harper, S. Lindley, and A. Sellen. Technology Heirlooms?: Considerations for Passing Down and Inheriting Digital Materials. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 337–346, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi:10.1145/2207676.2207723. → pages 15, 103, 133, 144

[213] W. Odom, A. Sellen, R. Harper, and E. Thereska. Lost in Translation: Understanding the Possession of Digital Things in the Cloud. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 781–790, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi:10.1145/2207676.2207789. → pages 2, 3, 16, 28, 131, 138, 139, 163

[214] W. Odom, J. Zimmerman, S. Davidoff, J. Forlizzi, A. K. Dey, and M. K. Lee. A Fieldwork of the Future with User Enactments. In *Proceedings of the Designing Interactive Systems Conference*, DIS '12, pages 338–347, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1210-3. doi:10.1145/2317956.2318008. → pages 20, 106, 115

[215] W. Odom, J. Zimmerman, and J. Forlizzi. Placelessness, Spacelessness, and Formlessness: Experiential Qualities of Virtual Possessions. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, DIS '14, pages 985–994, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2902-6. doi:10.1145/2598510.2598577. → pages 63, 65, 110, 131, 133, 139

[216] W. Odom, R. Wakkary, Y.-k. Lim, A. Desjardins, B. Hengeveld, and R. Banks. From Research Prototype to Research Product. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2549–2561, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858447. → pages 172

[217] W. Odom, R. Wakkary, J. Hol, B. Naus, P. Verburg, T. Amram, and A. Y. S. Chen. Investigating Slowness as a Frame to Design Longer-Term Experiences with Personal Data: A Field Study of Olly. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi:10.1145/3290605.3300264. → pages 15

[218] W. T. Odom, A. J. Sellen, R. Banks, D. S. Kirk, T. Regan, M. Selby, J. L. Forlizzi, and J. Zimmerman. Designing for Slowness, Anticipation and Re-visitation: A Long Term Field Study of the Photobox. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1961–1970, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi:10.1145/2556288.2557178. → pages 14, 15, 103

[219] K. E. Oh. Types of personal information categorization: Rigid, fuzzy, and flexible. *Journal of the Association for Information Science and Technology*, 68(6):1491–1504, 2017. → pages 72, 131, 133, 151

[220] K. E. Oh. Personal information organization in everyday life: modeling the process. *Journal of Documentation*, 75(3):667–691, 2019. → pages 72

[221] G. Oleksik, M. L. Wilson, C. Tashman, E. Mendes Rodrigues, G. Kazai, G. Smyth, N. Milic Frayling, and R. Jones. Lightweight Tagging Expands Information and Activity Management Practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 279–288, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi:10.1145/1518701.1518746. → pages 101, 102

[222] J. A. Oravec. Virtual Hoarding. In *Encyclopedia of Information Science and Technology, Fourth Edition*, pages 4306–4314. IGI Global, 2018. → pages 30, 174

[223] R. Orji, G. F. Tondello, and L. E. Nacke. Personalizing Persuasive Strategies in Gameful Systems to Gamification User Types. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi:10.1145/3173574.3174009. → pages 77, 96

[224] D. Orth, C. Thurgood, and E. V. D. Hoven. Designing Meaningful Products in the Digital Age: How Users Value Their Technological Possessions. *ACM Trans. Comput.-Hum. Interact.*, 26(5), Aug. 2019. ISSN 1073-0516. doi:10.1145/3341980. → pages 133

[225] L. Palen and P. Dourish. Unpacking "Privacy" for a Networked World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 129–136, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136307. doi:10.1145/642611.642635. → pages 133, 134

[226] S. T. Peesapati, V. Schwanda, J. Schultz, M. Lepage, S.-y. Jeong, and D. Cosley. Pensieve: Supporting Everyday Reminiscence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 2027–2036, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589299. doi:10.1145/1753326.1753635. → pages 15

[227] D. Petrelli and S. Whittaker. Family memories in the home: contrasting physical and digital mementos. *Personal and Ubiquitous Computing*, 14 (2):153–169, 2010. → pages 15, 16, 57

[228] D. Petrelli, S. Whittaker, and J. Brockmeier. AutoTopography: What Can Physical Mementos Tell Us about Digital Memories? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 53–62, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580111. doi:10.1145/1357054.1357065. → pages 15

[229] D. Petrelli, E. van den Hoven, and S. Whittaker. Making History: Intentional Capture of Future Memories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 1723–1732, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605582467. doi:10.1145/1518701.1518966. → pages 2, 14

[230] D. Petrelli, S. Bowen, and S. Whittaker. Photo mementos: Designing digital media to represent ourselves at home. *International Journal of Human-Computer Studies*, 72(3):320–336, 2014. → pages 15

[231] S. Petronio. *Boundaries of privacy: Dialectics of disclosure*. Suny Press, 2002. → pages 133, 134

[232] J. Pierce and E. Paulos. Counterfunctional Things: Exploring Possibilities in Designing Digital Limitations. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, DIS '14, page 375–384, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329026. doi:10.1145/2598510.2598522. → pages 133, 144

[233] E. Politou, E. Alepis, and C. Patsakis. Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. *Journal of Cybersecurity*, 4(1):tyy001, 2018. → pages 140

[234] A. Ponsard and J. McGrenere. Anchored Customization: Anchoring Settings to the Application Interface to Afford Customization. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4154–4165, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858129. → pages 219

[235] J. Porter. Google will soon let you auto-delete your location tracking data, May 2019. URL https://www.theverge.com/2019/5/1/18525384/google-location-tracking-data-auto-delete-history-app-and-activity-data-3-18-months. → pages 174

[236] C. Preist, D. Schien, and E. Blevis. Understanding and Mitigating the Effects of Device and Cloud Service Design Decisions on the Environmental Footprint of Digital Infrastructure. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1324–1337, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858378. → pages 2

[237] B. A. Price, K. Adam, and B. Nuseibeh. Keeping ubiquitous computing to yourself: A practical model for user control of privacy. *International Journal of Human-Computer Studies*, 63(1-2):228–253, 2005. → pages 102

[238] J. Pruitt and J. Grudin. Personas: Practice and Theory. In *Proceedings of the 2003 Conference on Designing for User Experiences*, DUX '03, pages 1–15, New York, NY, USA, 2003. ACM. ISBN 1-58113-728-1. doi:10.1145/997078.997089. → pages 77

[239] E. Rader. Yours, Mine and (Not) Ours: Social Influences on Group Information Repositories. In *Proceedings of the SIGCHI Conference on*

Human Factors in Computing Systems, CHI ’09, page 2095–2098, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605582467. doi:10.1145/1518701.1519019. → pages 12

[240] K. M. Ramokapane, A. Rashid, and J. Such. ”I feel stupid I can’t delete...”: a study of users’ cloud deletion practices and coping strategies. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, 2017. → pages 15, 17, 126

[241] P. Raschke, A. Küpper, O. Drozd, and S. Kirrane. Designing a GDPR-Compliant and Usable Privacy Dashboard. In *IFIP International Summer School on Privacy and Identity Management*, pages 221–236. Springer, 2017. → pages 140

[242] B. A. Richards and P. W. Frankland. The Persistence and Transience of Memory. *Neuron*, 94(6):1071–1084, 2017. → pages 49

[243] R. Rouge. Very Precious Memories: Digital Memories and Data Valorization. *Digitalization of Society and Socio-political Issues 1: Digital, Communication and Culture*, pages 71–79, 2019. → pages 95

[244] A. Rowstron and P. Druschel. Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. In R. Guerraoui, editor, *Middleware 2001*, pages 329–350, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-45518-9. → pages 164

[245] W.-J. Ryu, J.-H. Lee, K.-M. Kim, and S. Lee. MeCurate: Personalized Curation Service Using a Tiny Text Intelligence. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW ’17 Companion, page 269–272, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349147. doi:10.1145/3041021.3054723. → pages 132, 135

[246] A. Sahney and D. Loxton. Introducing Backup and Sync for Google Photos and Google Drive, July 2017. URL https://blog.google/products/photos/introducing-backup-and-sync-google-photos-and-google-drive/. → pages 140

[247] C. Sas and S. Whittaker. Design for Forgetting: Disposing of Digital Possessions After a Breakup. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’13, pages 1823–1832, New

York, NY, USA, 2013. ACM. ISBN 978-1-4503-1899-0. doi:10.1145/2470654.2466241. → pages 3, 16, 49, 63, 103

[248] C. Sas, S. Whittaker, S. Dow, J. Forlizzi, and J. Zimmerman. Generating Implications for Design through Design Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 1971–1980, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450324731. doi:10.1145/2556288.2557357. → pages 169

[249] C. Sas, S. Whittaker, and J. Zimmerman. Design for Rituals of Letting Go: An Embodiment Perspective on Disposal Practices Informed by Grief Therapy. *ACM Trans. Comput.-Hum. Interact.*, 23(4):21:1–21:37, Aug. 2016. ISSN 1073-0516. doi:10.1145/2926714. → pages 15, 64, 103, 126

[250] K. Schiele and M. Ucok Hughes. Possession rituals of the digital consumer: A study of Pinterest. *ACR European Advances*, 2013. → pages 30

[251] T. Schnitzler, C. Utz, F. M. Farke, C. Pöpper, and M. Dürmuth. Exploring user perceptions of deletion in mobile instant messaging applications. *Journal of Cybersecurity*, 6(1):tyz016, 2020. → pages 17

[252] S. Schoenebeck, N. B. Ellison, L. Blackwell, J. B. Bayer, and E. B. Falk. Playful Backstalking and Serious Impression Management: How Young Adults Reflect on Their Past Identities on Facebook. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1475–1487, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450335928. doi:10.1145/2818048.2819923. → pages 95

[253] D. Sedera and S. Lokuge. Is Digital Hoarding a Mental Disorder? Development of a Construct for Digital Hoarding for Future IS Research. In *Proceedings of the International Conference on Information Systems, ICIS 2018, San Francisco*, 2018. → pages 3, 16, 30, 174

[254] A. Sellen and S. Whittaker. Beyond total capture: a constructive critique of lifelogging. *Communications of the ACM*, 2010. → pages 2

[255] M. Seltzer and N. Murphy. Hierarchical File Systems Are Dead. In *Proceedings of the 12th Conference on Hot Topics in Operating Systems*, HotOS'09, page 1, USA, 2009. USENIX Association. → pages 10, 100

[256] D. Siegel and S. Dray. The map is not the territory: empathy in design. *interactions*, 26(2):82–85, 2019. → pages 77

[257] B. Signer. Towards Cross-Media Information Spaces and Architectures. In *2019 13th International Conference on Research Challenges in Information Science (RCIS)*, pages 1–7. IEEE, 2019. → pages 164

[258] V. V. Simbelis, P. Ferreira, E. Vaara, J. Laaksolahti, and K. Höök. Repurposing Bits and Pieces of the Digital. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 840–851, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi:10.1145/2858036.2858297. → pages 94

[259] D. Sinn, S. Kim, and S. Y. Syn. Personal digital archiving: influencing factors and challenges to practices. *Library Hi Tech*, 35(2):222–239, 2017. → pages 3, 69

[260] M. Sleeper, J. Cranshaw, P. G. Kelley, B. Ur, A. Acquisti, L. F. Cranor, and N. Sadeh. "i Read My Twitter the next Morning and Was Astonished": A Conversational Perspective on Twitter Regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 3277–3286, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi:10.1145/2470654.2466448. → pages 17

[261] F. Smadja. Automatic categorization of documents based on textual content, Sept. 16 2003. US Patent 6,621,930. → pages 160

[262] Spotify. Introducing Two New Personalized Playlists: On Repeat and Repeat Rewind, Sept. 2019. URL https://newsroom.spotify.com/2019-09-24/introducing-two-new-personalized-playlists-on-repeat-and-repeat-rewind/. → pages 95

[263] K. M. Spurgin. "Three backups is a minimum": A first look at norms and practices in the digital photo collections of serious photographers. *I, Digital: Personal Collections in the Digital Era, Chicago: Society of American Archivists*, pages 151–201, 2011. → pages 29, 30, 48

[264] N. Srnicek. *Platform capitalism*. John Wiley & Sons, 2016. → pages 2, 27, 49

[265] State of California. California Consumer Privacy Act (CCPA), Oct. 2018. URL https://oag.ca.gov/privacy/ccpa. → pages 1, 162, 173

[266] R. A. Stebbins. *Exploratory research in the social sciences*, volume 48. Sage, 2001. → pages 21

[267] A. Strauss and J. Corbin. Discovery of grounded theory. 1967. → pages 20

[268] T. Stylianou Lambert, L. A. Henkel, C. M. Weber, and K. Parisi. *Museums and Visitor Photography: Redefining the Visitor Experience*. Linda A. Henkel, Katelyn Parisi and Carey Mack Weber,"The Museum as Psychology Lab: Research on Photography and Memory in Museums," in Museums and Visitor Photography: Redefining the Visitor Experience, ed. Theopisti Stylianou-Lambert, Museumsetc, Edinburgh and Boston, 2016, pp. 152-83., 2016. → pages 49

[269] S. S. Sundar and S. S. Marathe. Personalization versus customization: The importance of agency, privacy, and power usage. *Human Communication Research*, 36(3):298–322, 2010. → pages 134

[270] L. Swan, A. S. Taylor, and R. Harper. Making place for clutter and other ideas of home. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 15(2):9, 2008. → pages 133

[271] G. Sweeten, E. Sillence, and N. Neave. Digital hoarding behaviours: Underlying motivations and potential negative consequences. *Computers in Human Behavior*, 85:54–60, 2018. → pages 3, 16, 30, 55, 119, 174

[272] B. Tarnoff. To decarbonize we must decomputerize: why we need a Luddite revolution. *The Guardian*, Sept. 2019. ISSN 0261-3077. URL https://www.theguardian.com/technology/2019/sep/17/tech-climate-change-luddites-data. → pages 173

[273] L. Tauscher and S. Greenberg. How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, 47(1):97–137, 1997. → pages 10

[274] J. Teevan, R. Capra, and M. P. Quiñones. How people find personal information. *Jones et Teevan (2007b), chapitre*, 2:22–34, 2007. → pages 73

[275] Y.-Y. Teing, A. Dehghantanha, K.-K. R. Choo, T. Dargahi, and M. Conti. Forensic investigation of cooperative storage cloud service: Symform as a case study. *Journal of forensic sciences*, 62(3):641–654, 2017. → pages 164

[276] The Economist. The world's most valuable resource is no longer oil, but data, May 2017. URL https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource. → pages 27

[277] The Guardian. The NSA files, 2018. URL https://www.theguardian.com/us-news/the-nsa-files. → pages 4

[278] A. Thudt, D. Baur, S. Huron, and S. Carpendale. Visual mementos: Reflecting memories with personal data. *IEEE transactions on visualization and computer graphics*, 22(1):369–378, 2015. → pages 15

[279] G. F. Tondello, R. R. Wehbe, L. Diamond, M. Busch, A. Marczewski, and L. E. Nacke. The Gamification User Types Hexad Scale. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '16, page 229–243, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344562. doi:10.1145/2967934.2968082. → pages 77, 96

[280] S. J. Tracy. Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative inquiry*, 16(10):837–851, 2010. → pages 23, 96

[281] T. A. Tran, S. Schwarz, C. Niederée, H. Maus, and N. Kanhabua. The Forgotten Needle in My Collections: Task-Aware Ranking of Documents in Semantic Information Space. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 13–22, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3751-9. doi:10.1145/2854946.2854971. → pages 101, 102

[282] A. Ukil, S. Bandyopadhyay, and A. Pal. Privacy for IoT: Involuntary privacy enablement for smart energy systems. In *2015 IEEE International Conference on Communications (ICC)*, pages 536–541, 2015. → pages 155

[283] A. Valdez. Everything You Need to Know About Facebook and Cambridge Analytica — WIRED, Mar 2018. URL https://www.wired.com/story/wired-facebook-cambridge-analytica-coverage/. → pages 4

[284] M. J. van Bennekom, R. M. Blom, N. Vulink, and D. Denys. A case of digital hoarding. *BMJ case reports*, 2015:bcr2015210814, 2015. → pages 30

[285] G.-J. van Bussel, N. Smit, and J. van de Pas. Digital archiving, green IT and environment. Deleting data to manage critical effects of the data deluge. *Electronic Journal of Information Systems Evaluation*, 18(2):187, 2015. → pages 2

[286] M. Van Kleek and K. OHara. The future of social is personal: The potential of the personal data store. In *Social Collective Intelligence*, pages 125–158. Springer, 2014. → pages 102

[287] M. Van Kleek, D. A. Smith, D. Murray Rust, A. Guy, K. O'Hara, L. Dragan, and N. R. Shadbolt. Social Personal Data Stores: The Nuclei of Decentralised Social Machines. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 1155–1160, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334730. doi:10.1145/2740908.2743975. → pages 164

[288] J. Vertesi, J. Kaye, S. N. Jarosewski, V. D. Khovanskaya, and J. Song. Data Narratives: Uncovering Tensions in Personal Data Management. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 478–490, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3592-8. doi:10.1145/2818048.2820017. → pages 10, 11, 14, 16, 28, 32, 35, 41, 90, 91, 96, 117, 123, 131, 133, 140, 151, 153, 159, 163

[289] F. Vitale. Designing for Long-Term Digital Data Management. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359719. doi:10.1145/3290607.3299070. → pages viii

[290] F. Vitale, I. Janzen, and J. McGrenere. Hoarding and Minimalism: Tendencies in Digital Data Preservation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 587:1–587:12, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3174161. → pages vi, 27

[291] F. Vitale, W. Odom, and J. McGrenere. Keeping and Discarding Personal Data: Exploring a Design Space. In *Proceedings of the 2019 on Designing

Interactive Systems Conference, DIS '19, page 1463–1477, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5850-7. doi:10.1145/3322276.3322300. → pages vii, 98

[292] F. Vitale, J. Chen, W. Odom, and J. McGrenere. Data Dashboard: Exploring Centralization and Customization in Personal Data Curation. In *Proceedings of the 2020 ACM on Designing Interactive Systems Conference*, DIS '20, page 311–326, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369749. doi:10.1145/3357236.3395457. → pages viii, 130

[293] A. Voida, J. S. Olson, and G. M. Olson. Turbulence in the Clouds: Challenges of Cloud-based Information Work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2273–2282, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1899-0. doi:10.1145/2470654.2481313. → pages 12, 42, 47, 133

[294] S. Voida and S. Greenberg. WikiFolders: Augmenting the Display of Folders to Better Convey the Meaning of Files. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1679–1682, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi:10.1145/1518701.1518959. → pages 100, 101, 102

[295] S. Voida and E. D. Mynatt. It Feels Better Than Filing: Everyday Work Experiences in an Activity-based Computing System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 259–268, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi:10.1145/1518701.1518744. → pages 101, 102, 133

[296] S. Voida, W. K. Edwards, M. W. Newman, R. E. Grinter, and N. Ducheneaut. Share and Share Alike: Exploring the User Interface Affordances of File Sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 221–230, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933727. doi:10.1145/1124772.1124806. → pages 133

[297] D. Vorick and L. Champine. Sia: Simple decentralized storage. *Nebulous Inc*, 2014. → pages 164

[298] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. "I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450309110. doi:10.1145/2078827.2078841. → pages 17

[299] R. D. Watkins, A. Sellen, and S. E. Lindley. Digital Collections and Digital Collecting Practices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3423–3432, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi:10.1145/2702123.2702380. → pages 14, 65, 127, 133

[300] M. Weiland and R. Dachselt. Facet Folders: Flexible Filter Hierarchies with Faceted Metadata. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 3735–3740, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-012-8. doi:10.1145/1358628.1358922. → pages 100, 101, 102

[301] R. Whitham and L. Cruickshank. The Function and Future of the Folder. *Interacting with Computers*, 29(5):629–647, 2017. doi:10.1093/iwc/iww042. → pages 100, 133

[302] S. Whittaker. Personal information management: from information consumption to curation. *Annual review of information science and technology*, 45(1):1–62, 2011. → pages 2, 3, 12, 13, 15, 28, 29, 72, 73

[303] S. Whittaker and J. Hirschberg. The Character, Value, and Management of Personal Paper Archives. *ACM Trans. Comput.-Hum. Interact.*, 8(2): 150–170, June 2001. ISSN 1073-0516. doi:10.1145/376929.376932. → pages 10

[304] S. Whittaker and C. Sidner. Email Overload: Exploring Personal Information Management of Email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 276–283, New York, NY, USA, 1996. ACM. ISBN 0-89791-777-4. doi:10.1145/238386.238530. → pages 102

[305] S. Whittaker, Q. Jones, and L. Terveen. Contact Management: Identifying Contacts to Support Long-term Communication. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pages 216–225, New York, NY, USA, 2002. ACM. ISBN 1-58113-560-2. doi:10.1145/587078.587109. → pages 10

[306] S. Whittaker, V. Bellotti, and J. Gwizdka. Email in Personal Information Management. *Commun. ACM*, 49(1):68–73, Jan. 2006. ISSN 0001-0782. doi:10.1145/1107458.1107494. → pages 10, 72

[307] S. Whittaker, O. Bergman, and P. Clough. Easy on that trigger dad: a study of long term family photo retrieval. *Personal and Ubiquitous Computing*, 14(1):31–43, 2010. → pages 14, 16

[308] S. Whittaker, V. Kalnikaitė, D. Petrelli, A. Sellen, N. Villar, O. Bergman, P. Clough, and J. Brockmeier. Socio-Technical Lifelogging: Deriving Design Principles for a Future Proof Digital Past. *Human-Computer Interaction*, 27(1-2):37–62, 2012. doi:10.1080/07370024.2012.656071. URL https://www.tandfonline.com/doi/abs/10.1080/07370024.2012.656071. → pages 3, 11

[309] P. Williams, J. Leighton John, and I. Rowland. The personal curation of digital objects: A lifecycle approach. In *Aslib Proceedings*, volume 61, pages 340–363. Emerald Group Publishing Limited, 2009. → pages 4, 70

[310] J. O. Wobbrock and J. A. Kientz. Research Contributions in Human-computer Interaction. *Interactions*, 23(3):38–44, Apr. 2016. ISSN 1072-5520. doi:10.1145/2907069. → pages 5

[311] M. K. Wolters, E. Niven, M. Runardotter, F. Gallo, H. Maus, and R. H. Logie. Personal Photo Preservation for the Smartphone Generation. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1549–1554, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3146-3. doi:10.1145/2702613.2732793. → pages 62

[312] R. Y. Wong, D. K. Mulligan, E. Van Wyk, J. Pierce, and J. Chuang. Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):111:1–111:26, Dec. 2017. ISSN 2573-0142. doi:10.1145/3134746. → pages 128

[313] World Backup Day. World Backup Day, n.d. URL http://www.worldbackupday.com. → pages 94

[314] C. Zeidler, C. Lutteroth, and G. Weber. An Evaluation of Advanced User Interface Customization. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, OzCHI '13, page 295–304, New York, NY,

USA, 2013. Association for Computing Machinery. ISBN 9781450325257. doi:10.1145/2541016.2541037. → pages 134

[315] X. Zhao and S. E. Lindley. Curation Through Use: Understanding the Personal Value of Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2431–2440, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi:10.1145/2556288.2557291. → pages 14, 127

[316] J. Zimmerman. Video Sketches: Exploring pervasive computing interaction designs. *IEEE pervasive computing*, 4(4):91–94, 2005. → pages 99, 106

[317] J. Zimmerman, J. Forlizzi, and S. Evenson. Research Through Design As a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 493–502, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9. doi:10.1145/1240624.1240704. → pages 20, 73, 99, 104, 133, 144

[318] J. Zimmerman, E. Stolterman, and J. Forlizzi. An Analysis and Critique of Research through Design: Towards a Formalization of a Research Approach. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, page 310–319, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450301039. doi:10.1145/1858171.1858228. → pages 20

[319] S. Zuboff. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1):75–89, 2015. → pages 2, 162

[320] X. Zürn, M. Broekhuijsen, D. van Gennip, S. Bakker, A. Zijlema, and E. van den Hoven. Stimulating Photo Curation on Smartphones. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 255–259, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450360258. doi:10.1145/3295750.3298947. → pages 94, 132

Appendix A

Data Dashboard Usability Report

Below, we include the complete usability report for Data Dashboard. It contains more details about how participants used different sections and functions during the scenarios, together with suggested design changes. I wrote the report together with undergraduate student Janet Chen.

A.1 User interface usability

A.1.1 Activity

Most [1] participants used the Activity section at some point during the scenarios. However, Explore Your Data and Quick Actions were more popular [2]. When asked about it, several participants said they found Activity underwhelming or frustrating. They thought it was unhelpful because it was too similar to Explore Your Data, but not as complete. The option of filtering data by time appeared generally useful but a few participants found it confusing.

Suggested changes

- Consider merging Activity and Explore Your Data.

[1] Activity used at least once by 10/18 participants.

[2] Explore Your Data: used by 17/18 participants at least once, used 40 times across all participants; Quick Actions: used by 15/18 participants at least once, used 33 times across all participants.

- Adding a time filtering option to Explore Your Data should provide all the functions of Activity with none of the redundancy.

A.1.2 Shared Data

Within Activity, most participants did not notice the Shared Data section. When prompted about it, some thought it could be helpful, but several misunderstood how it works (they thought other people were seeing all of the data in the system). Overall, several participants questioned the need or usefulness of Shared Data for managing their own personal items.

Suggested changes

- Consider removing Shared Data.
- Details about sharing could be included on individual items in other sections.

A.1.3 Explore Your Data

Most participants used Explore Your Data (EYD) at some point in the scenarios. In fact, it was the most commonly used section throughout the scenarios. Participants thought that EYD was useful to get an overview of all personal data, a bird-eye perspective, as one participant said. They also thought that EYD would be better suited for occasional, more focused scenarios, as opposed to short management episodes, where Quick Actions would work better.

Data types

In general, participants thought that organizing data automatically into different types and lists based on different filtering criteria (i.e., data to get rid of, data to organize, private information, data to back up or sync) was good. But a few participants said they do not think about their data in terms of types largely based on formats and file extensions. Instead they think in terms of projects organized into folders. A few participants wanted to see all data categorized into lists, not just the items that fit into the sidebar filters.

Several participants wanted to customize the list of data types in EYD, both in terms of content and order. Some wanted to add their own data types. Only half of the participants discovered the button to customize the list of data types on their own ("Sort or hide the types of data you see"). But when prompted to use it, almost all participants found it very helpful. Most data types listed in EYD seemed to resonate with participants, except Tracking Data and Cache & Logs. Some participants did not know what Tracking Data or Cache could be.

Lists of items

When considering the lists of items in EYD, some participants wanted an indication of data recency. Some participants also found the amount of text in EYD's lists overwhelming.

When clicking on "Data I could get rid of" in the sidebar filters, the indication of space taken by different data items was useful, but participants also wanted an indication of the total remaining and occupied space on different devices and platforms.

Suggested changes

- Add filtering and sorting by time and platform/device.
- Show total free/occupied space for different types and platforms/devices.
- Make "sort and hide" work direct manipulation rather than a popup panel and increase its discoverability.
- Consider categorizing all data not just data that matches the filters.
- Better show on which device/platform items are (also in QA).
- (See below for specific design changes for sidebar filters.)

A.1.4 Quick Actions

Participants used Quick Actions (QA) frequently during the scenarios. They thought it was helpful for freeing up space and organizing data regularly, in short amounts of time. Most participants expected a machine learning algorithm to be behind the suggestions and they expected QA to focus on showing recent or "on-going" items. Some participants felt that seeing actual data was more helpful than having lists of links like in EYD. Only a few participants did not find QA useful and said that it felt too restrictive compared to EYD.

In terms of specific UI elements, several participants did not use the filters in QA and instead scrolled through the suggestions. Several participants found the "snoozing" button for each suggestion confusing.

Suggested changes

- Remove the snoozing button.
- Show on which device/platform an item is stored more consistently and more prominently.
- Show a preview of the actual data if possible.
- Make it possible to mark exceptions in the suggested actions (as in, do not suggest this item again).
- Consider removing the sidebar filters from QA.

A.1.5 Sidebar filter panel

The sidebar filter panel is both in EYD and QA. All participants except one discovered the filters on their own, and most used them at some point during the scenarios. When discussing them in more detail, most participants found the filters comprehensive and the idea of automatically clustering data helpful. However, several participants found the filters confusing or to be "too much work". A few thought some of the labels should be renamed or were unclear (the word "sync" in particular was not clear to a few participants).

Suggested changes

- Think of a different UI pattern for the filters.

- Consider changing the labels for different (e.g., “declutter” or “free up space,” “organize,” “back up”).

A.1.6 Settings

Most participants discovered Settings and its link to the sidebar filters in EYD and QA. However, roughly half of the participants did not fully understand the connection between Settings and the filters. In general, several participants found the Settings page confusing: for example, they expected a button to make the Settings “run” or they expected that after checking one of the options the system would start deleting or organizing data immediately. Eventually, after being prompted about them, some of these participants understood the link between Settings and the sidebar filters. Participants who understood the Settings-filters mechanism thought that customizing filters was a good idea. Settings made participants feel in charge of the system. Several thought that the default options were comprehensive and novel. Being able to create additional combinations was confusing for some participants but generally seen as helpful.

Suggested changes

- Emphasize the link between filters and settings by anchoring [234] the advanced settings to the filters in EYD and QA.
- Expand the list of defaults to cover all major cases and situations.
- Consider a quicker way to include new combinations, maybe through an online repository of filters that people can add and share.

A.2 Use scenarios

A.2.1 Scenario 1

Scenario 1 (*space running out triggers curating actions*): “The space on your computer is running out. You want to find some data to discard. You are not sure where to start looking, but you know that you do not care too much about old documents.”

In the first scenario, most participants used EYD to find documents to discard. Often participants tried to click “browse all” in each data type to manually explore their data. Some participants used the sidebar filter “Data to get rid of” and appreciated what the system suggested. Several participants also used the suggested actions in QA to remove duplicates. However, a few participants tried to use the Settings page to “delete” data from one of the default options (e.g. “Data older than 5 years”) and found this function confusing.

Implications: EYD and QA seem to fit expectations and filters seem helpful for going through different scenarios. Settings can be confusing because of the issues mentioned before.

A.2.2 Scenario 2

Scenario 2 (*taking time for regular data curation*): “It is a rainy day. You have set aside some time for doing a regular cleanup of your devices. You usually do this every few months. You want to review your data and make sure everything is organized in your preferred way.”

When organizing in the preferred way, the majority of participants used EYD with many selecting the “Data to organize” filter. They often listed their preferred order to look through data types (e.g., Messages, Photos, Emails, etc.). The order changed based on personal preference, but most participants wanted to go through EYD cards one by one. Several participants wanted to sort by data type and recency so they used Activity instead of EYD. For example, one participant went to Activity to look for unorganized recent activity to move into folders.

Implications: once again, filters seem helpful, and EYD meets expectations.

A.2.3 Scenario 3

Scenario 3 (*exploring recent data*): “You have 5 to 10 minutes in between meetings and errands. You decide to take a look at your recent data to get a sense of anything that needs taking care of.”

Almost half of the participants used QA to squeeze in some data management within 5 to 10 minutes. In QA, most participants used the options to rename files, backup and sync, remove duplicates, and move folders into files. Other participants preferred to browse through all their data with EYD or use Activity to check for unread documents or unopened emails. One participant wanted to use Shared Data "to see if anyone [had] done anything on the documents". This suggests that Shared Data could help brief check-ins on collaborative projects during dead moments.

Implications: recommendations seem to fit the urgency mental model. Shared Data likely has more potential as part of collaboration-focused systems.

A.2.4 Scenario 4

Scenario 4 (*protecting data privacy*): "You have heard about a data leak from a popular cloud storage platform that exposed personal information to hackers. You want to review what data you have stored on different cloud platforms that might pose a privacy risk in the future."

When going through the privacy scenario, several participants scanned the system looking for a way to see all data stored in a specific platform (e.g., iCloud, Dropbox, Google Drive) or device (e.g., "my drive"). Often they could not find what they were looking for. For example, one participant went to the Settings section and then to Accounts at the bottom, hoping they would be able to access all data on Dropbox from the link to the Dropbox account. Another scrolled through Quick Actions until they saw the Dropbox and iCloud icons.

Implications: participants looked for visual *signifiers* (i.e., icons) of platforms and devices where the data is stored. With privacy scenarios, their mental model focuses on specific platforms rather than specific data types. A possible change would be to make it possible to see data grouped by platform as an alternative to the current grouping by data types.

A.2.5 Scenario 5

Scenario 5 (safeguarding data spread across devices and platforms): "You are in the process of buying a new computer. You want to make sure that you are not going to lose any of the data you care about. You want to ensure that everything is stored in more than one place."

Many participants were unable to use the system to go through the last scenario. Some participants used EYD to access all files but were left wondering if there was an option to back them up to the cloud or save to an external hard drive. Several participants talked about their actual process of backing up files and did not use the system. But some participants did use the "backup and sync" filter in QA.

Implications: backing up seems to be either a completely automatic process (with Time Machine and Google Photos) that requires little user input and where all items are backed up, or a completely manual process with external hard drives. This might be the least relevant action and given the confusion around the term "sync" this filter might be demoted or redesigned. A possible change would be to make the "Back up" section focus more on exporting data or maybe providing a visualization of backed up items.

Appendix B

Survey Questions and Additional Results

B.1 Complete set of survey questions

Q4 We are going to ask you questions about your ***digital data***.

By **digital data**, we mean anything that you consider your own "stuff" on the digital devices you own: for example, your computer files, your emails, your photos, your texts, messages (e.g., WhatsApp conversations), or mobile apps.

Q5 What is your general approach in deciding what digital data to keep or delete over the long term?

Please choose where you fall on the spectrum.

	I delete as much as possible	I delete more than I keep	Somewhere in the middle	I keep more than I delete	I keep as much as possible
General approach in keeping or deleting data	○	○	○	○	○

Q6 How does your approach differ for each of the following data types?

	I delete as much as possible	I delete more than I keep	Somewhere in the middle	I keep more than I delete	I keep as much as possible	❌ (Does not apply)
Documents on one of your devices (e.g., Word, Excel, pdf files)	○	○	○	○	○	○
Documents on a cloud storage platform (e.g., Google Docs)	○	○	○	○	○	○
Texts and messages	○	○	○	○	○	○
Media files (e.g., video, audio files)	○	○	○	○	○	○
Pictures	○	○	○	○	○	○
Emails	○	○	○	○	○	○
Browser bookmarks	○	○	○	○	○	○
Mobile applications	○	○	○	○	○	○
Contacts	○	○	○	○	○	○
Facebook friends	○	○	○	○	○	○

Q7 Choose how the following statements reflect your experience

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I am satisfied with my general approach for selecting what data to keep or delete	○	○	○	○	○	○	○
I can always find the data I am looking for	○	○	○	○	○	○	○
I take steps to prevent losing my data (e.g., backup)	○	○	○	○	○	○	○
I spend a lot of time managing my data	○	○	○	○	○	○	○
I think selecting what data to keep or delete is a good use of my time	○	○	○	○	○	○	○
My data represent important parts of my life	○	○	○	○	○	○	○
I do not like spending a lot of time using interactive devices (e.g.,	○	○	○	○	○	○	○

computer, smartphone, tablet)							
I have a clear idea of what data I have across my platforms and devices	○	○	○	○	○	○	○
I am ok with the possibility of losing some of my data	○	○	○	○	○	○	○
I do not like to let go of things from my past	○	○	○	○	○	○	○
I am not particularly attached to my data	○	○	○	○	○	○	○
I think most people have a similar approach to mine	○	○	○	○	○	○	○
Deciding what data to delete takes too much time	○	○	○	○	○	○	○
I find it hard to manage my data with current tools	○	○	○	○	○	○	○
I have adequate tools to select what data to keep or delete	○	○	○	○	○	○	○

I would like to delete some of my data that I do not need anymore	○	○	○	○	○	○	○
I have clutter in my data that I would like to take care of	○	○	○	○	○	○	○
I know I have some data that I do not need anymore but I am not sure where it is	○	○	○	○	○	○	○
I acquire or create large amounts of data on a regular basis	○	○	○	○	○	○	○

Q8 When was the **last time you decluttered some of your digital data**, that is, you were in an active session focused mostly on deleting some of your data?

○ Never

○ More than a year ago

○ Within the last year

○ Within the last 6 months

○ Within the last month

○ Within the last week

Q9 Can you briefly describe what you decluttered?

You can touch on these aspects: What data did you delete? What prompted you to do it? What was your process? What was challenging? What was the final result?

____

____

____

____

____

Q10 How often do you declutter digital data?

- Never
- Rarely (e.g., when running out of space, when buying a new device)
- Once a year
- Multiple times a year
- Monthly
- Weekly
- Daily

Q11 What is your general approach in **organising digital data**?

Please take a quick look at your data as it is now and choose where you fall on a spectrum between unstructured (that is, you do not use folders and subfolders, tags or labels; you do not rename files) and **structured** (that is, you have a structure with folders and subfolders, or tags and labels; you rename files).

	Extremely unstructured	Unstructured	Somewhere in between	Structured	Extremely structured	❌ (Does not apply)
Documents on a device	◯	◯	◯	◯	◯	◯
Cloud documents	◯	◯	◯	◯	◯	◯
Media files	◯	◯	◯	◯	◯	◯
Pictures	◯	◯	◯	◯	◯	◯
Emails	◯	◯	◯	◯	◯	◯
Browser bookmarks	◯	◯	◯	◯	◯	◯
Mobile applications	◯	◯	◯	◯	◯	◯
Contacts	◯	◯	◯	◯	◯	◯

End of Block: Digital data

Start of Block: Devices

Q12 Now we are going to ask you questions about your **digital devices** (e.g., computer, smartphone).

Q13 How many interactive devices (e.g., laptop, smartphone, tablet, smartwatch) do you regularly use?

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ more than 3

Q14 How much storage space does your main computer provide?

- ◯ 128GB or less
- ◯ Up to 256GB
- ◯ Up to 512GB
- ◯ Up to 1TB
- ◯ More than 1TB
- ◯ I am not sure where to look this up
- ◯ I do not own or regularly use a computer

Q15 How much space is available on your computer right now?

- ◯ Less than half of the total available space
- ◯ About half of the total available space
- ◯ More than half of the total available space
- ◯ I am not sure where to look this up

Q16 How long have you had your main computer for?

- ◯ Less than a year
- ◯ Between 1 and 5 years
- ◯ More than 5 years
- ◯ I don't remember

Q17 How much storage space does your smartphone provide?

- ◯ 16GB or less
- ◯ Up to 64GB
- ◯ Up to 256GB
- ◯ I am not sure where to look this up
- ◯ I do not own or regularly use a smartphone

Q18 How much space is available on your smartphone right now?

- ◯ Less than half of the total available space
- ◯ About half of the total available space
- ◯ More than half of the total available space
- ◯ I am not sure where to look this up

Q19 How long have you had your smartphone for?

- ◯ Less than 1 year
- ◯ Between 1 and 3 years
- ◯ More than 3 years
- ◯ I don't remember

End of Block: Devices

Start of Block: Additional information

Q20
You are almost done, thanks for sticking with us!

Q21 We are now switching gears and are asking about your **physical possessions** (rather than digital ones) in your home (across different rooms) and in your physical workspace at your place of work/study.

Where does your organisation fall between **messy** and **tidy**?

	Definitely messy	More messy than tidy	In between messy and tidy	More tidy than messy	Definitely tidy	❌ (Does not apply)
Bedroom	◯	◯	◯	◯	◯	◯
Closet	◯	◯	◯	◯	◯	◯
Kitchen area	◯	◯	◯	◯	◯	◯
Bathroom	◯	◯	◯	◯	◯	◯
Work / study area	◯	◯	◯	◯	◯	◯

Q22 When was the last time you **decluttered some of your physical possessions**?

- ◯ Never
- ◯ More than a year ago
- ◯ Within the last year
- ◯ Within the last 6 months
- ◯ Within the last month
- ◯ Within the last week

Q23 Can you briefly describe what prompted you to declutter your possessions?

____

____

____

____

____

Q24 How often do you declutter physical possessions?

- ○ Never
- ○ Rarely (e.g., when moving to a new place)
- ○ Once a year
- ○ Multiple times a year
- ○ Monthly
- ○ Weekly
- ○ Daily

Q25 What is your current occupation?

____

Q26 Choose how the following statements reflect your experience

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disagree	Strongly disagree
My days are usually full of activities at work, home, school, or elsewhere	○	○	○	○	○	○	○
I value leading a busy life	○	○	○	○	○	○	○
I want to use my time as efficiently as possible	○	○	○	○	○	○	○
I regularly have some free time to unwind or relax	○	○	○	○	○	○	○
I feel like I always have a lot to do and not enough time	○	○	○	○	○	○	○

Q27 What is your cultural background?
(e.g., Native American, North American, Southern Asian, etc.)

____

Q28 What is the highest level of education you have completed?

- None
- Primary or high school
- Technical training
- Bachelor level
- Master level
- PhD level

Q29 Do you have a technical background (e.g., Computer Science) or experience with programming?

- Yes
- No

Q30 How old are you?

- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65-74
- 75 and above

Q31 What term best describes your gender identity?

- ○ Male
- ○ Female
- ○ Transgender
- ○ Other
- ○ Prefer not to answer

End of Block: Additional information

B.2 Descriptive results about general data practices

In their general approach, a slight majority of survey respondents (51%) tended towards keeping over deleting data (combined bottom two rows in Table B.1).

Approach	Percentage	Count
I delete as much as possible	8.60%	30
I delete more than I keep	12.89%	45
Somewhere in the middle	27.79%	97
I keep more than I delete	36.68%	128
I keep as much as possible	14.04%	49
Total		349

Table B.1: In their general approach, survey participants tended towards keeping over deleting digital data.

When it came to specific data types (Figure B.1), respondents generally reported more keeping for pictures (74%, combining the counts for the two options "keeping most" and "keeping more than deleting" together), contacts (71%), and Facebook friends (59%). Instead, they reported more deleting for mobile applications (42%, similarly counting the two options "deleting most" and "deleting more than keeping" together), emails (35%), and bookmarks (30%). These results support the idea that keeping and deleting decisions can vary across data types. However, the limitations in our sample and phrasing of these questions prevent us from making broader claims about these results.

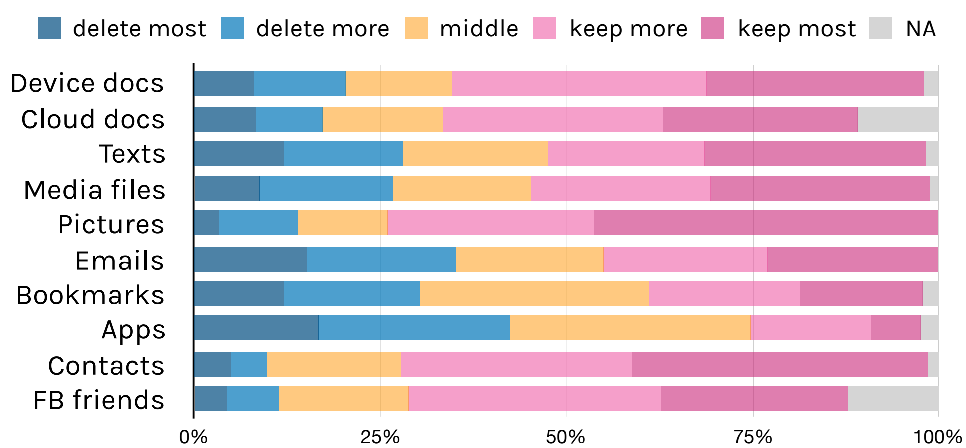


Figure B.1: Keeping and deleting decisions varied according to different data types: respondents reported the most keeping for pictures and the most deleting for mobile apps.

Appendix C

Data Analysis

C.1 Examples of coding in Study 1

In Study 1, we used Saturate[1] to conduct collaborative coding of the interviews. Here, two examples of coding that show some categories paired with codes (Figure C.1) and the collaborative nature of the process (Figure C.2).

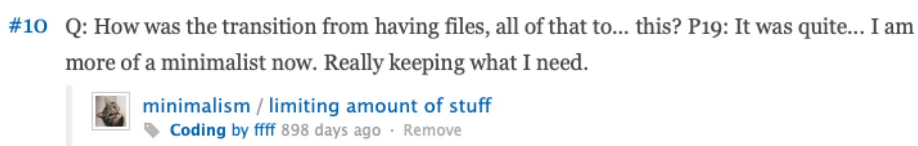


Figure C.1: Example of a category and related code from Study 1.

[1] http://www.saturateapp.com

#31 Q: What do you think will happen to your digital data one year from now? P2: I don't think anything is going to go. I'm just going to add more because it costs so little to add stuff but it takes a lot of time to sort the stuff you want to delete.

hoarding / low cost
Coding by Izabelle Janzen 889 days ago · Remove

hoarding / avoiding to delete data
Coding by ffff 906 days ago · Remove

I might have a code here about hoarding due to low cost. my sense is for alot of people data is hoarded because there is little reason not to. That it's relatively easy to ...
Memo by Izabelle Janzen 889 days ago · sentence 4

Figure C.2: Example of collaborative coding from Study 1.

Category	Codes in category
backup 24 codes applied to 74 paragraphs	barriers 6x confusion around format of backups 1x how – automatic 6x how – automatic -- confusing 6x how – automatic + manual because of quality issues 1x how – automatic -- satisfaction 1x how – cloud 3x how – duplicate on same device 2x how – email 5x how – external hard drive 5x 14 more →
cloud vs. hardware 12 codes applied to 49 paragraphs	cloud as a forced choice 1x cloud > hardware 4x distrust in cloud 6x distrust in hardware 17x general distrust 2x hardware > cloud 2x physicality of hardware 6x trust in a brand 2x trust in hardware 2x trust in the cloud 14x 2 more →

Figure C.3: Example of categories and codes from Study 1.

hoarding 29 codes applied to 79 paragraphs	avoiding cloud to curb the amount of stuff 1x avoiding to delete data 17x cleaning up is only a distant possibility 1x collecting data 3x confusion about deleting data 2x data accumulates 1x data is essential in life 1x deleting only because of space 1x difficulty of letting things go 1x does all that stuff bring value? 1x 19 more →

Figure C.4: Examples of codes about “hoarding” in Study 1.

minimalism 18 codes applied to 45 paragraphs	basic functionality 2x being content 3x cleaning up 10x face to face interaction more important than technology 2x feeling calm inside 1x keeping things clean 1x life is more than your data 2x limiting amount of stuff 9x minimal approach to life 2x not being techy 6x 8 more →

Figure C.5: Examples of codes about “minimalism” in Study 1.

C.2 Memos from Study 2

These memos contain rough and preliminary ideas. I reproduce them here in their original form, but they might present concepts that significantly evolved over time.

April 6, 2018 (Study 2 ongoing):

> different dimensions or types of minimalism: aesthetic minimalism, minimal technology use, minimal data storage, minimal possessions, minimalism as a life philosophy

May 8, 2018 (Study 2 ongoing):

> the way people approached things (physical or digital or both) in their process was both retrospective (e.g., I haven't used in a long time, I can get rid of it) and prospective (e.g., this will likely going to cause me difficulties, better get rid of it now). So looking back, but also looking forward – things (and the self) through time

May 22, 2018 (Study 2 ongoing):

> Decluttering involves selection and deletion
>
> Declutter: removing unwanted items
>
> Select: choose a subset from a larger group
>
> Delete: remove
>
> Maybe, it makes sense to see it this way: decluttering is the process, deleting and selecting are the actions.

June 2, 2018 (Study 2 ongoing):

> Long term solutions won't come from the file and folders paradigm
>
> A comprehensive view of all can be helpful, maybe a dashboard with metadata to highlight things that are likely candidates to be preserved (for example the files, or better the objects, you spent the most time on) or deleted (the files you never opened or haven't opened in x etc.)

June 18, 2018 (Study 2 ongoing)

when do people delete data?

The *when*:

regularly (i.e., regular time they set to go through and delete their data, short independent episodes over longer periods of time – this is *routine decluttering* – the goal is to delete data as an activity in itself)

when an unusual event happens (e.g., breakup, new device, space is running low – this is *triggered decluttering*)

in the context of a related but different daily activity (e.g., listening to songs, interacting with data by doing other actions where the main goal is not deleting something – *contextual decluttering* – actually accidental decluttering better describes it)

C.3 Example of coding in Study 3

In Study 3 I used Annotations, a Mac application[2] to analyse interview transcripts. Here (Figure C.6), an example of codes that became part of the theme "selecting data is a personal responsibility": "automation makes you look lazy" and "I can think on my own."

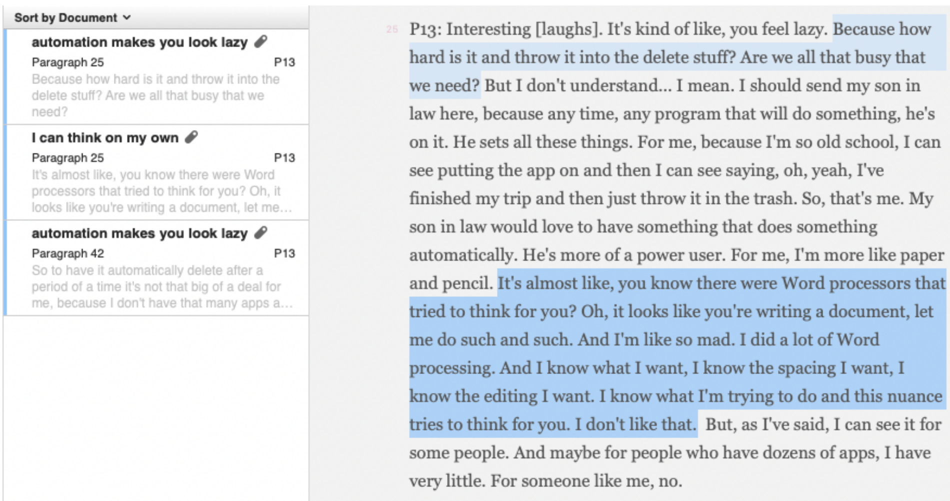


Figure C.6: Example of coding from Study 3.

[2] https://www.annotationsapp.com

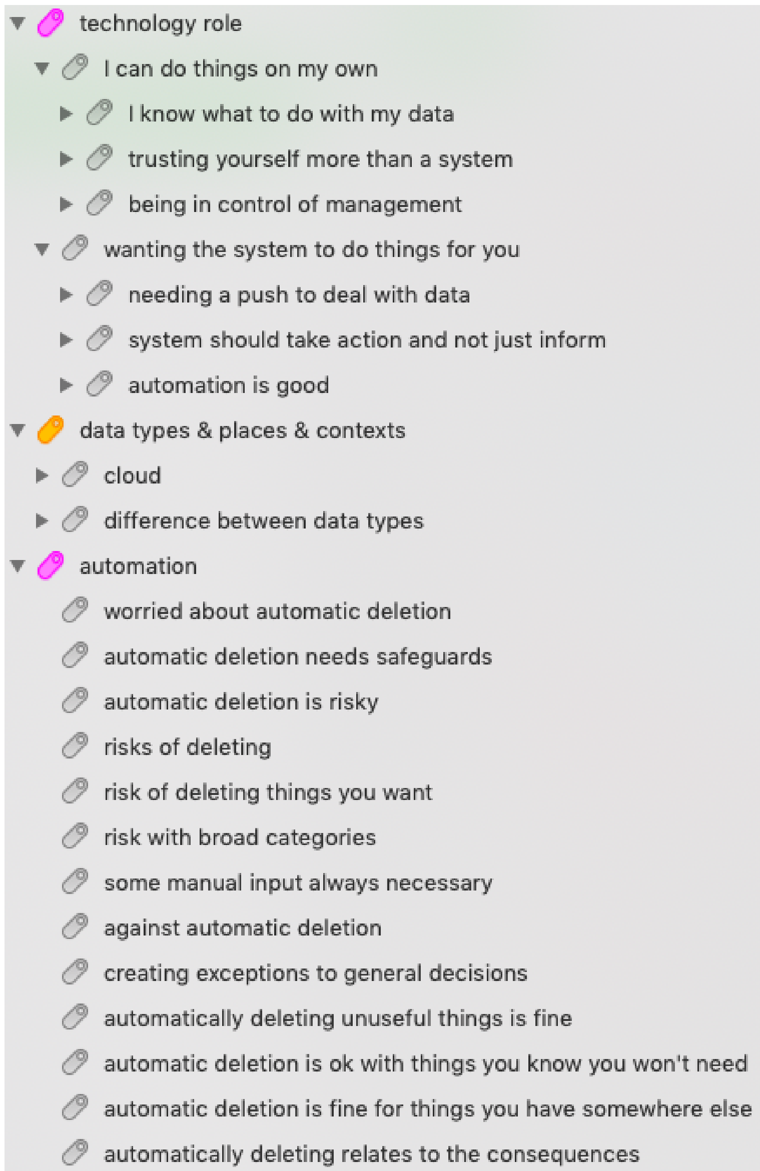
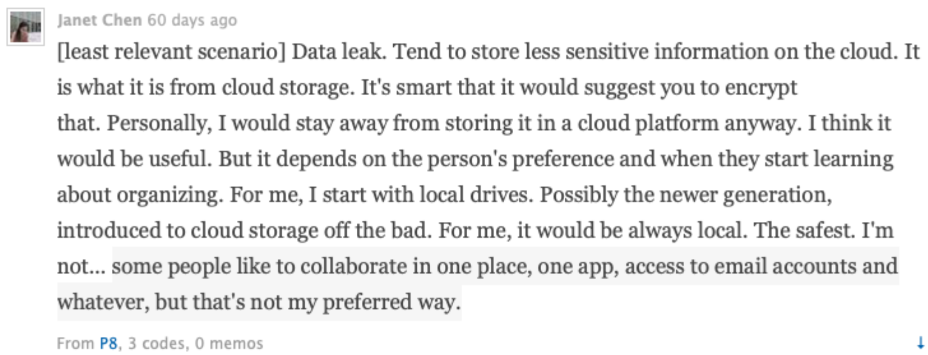


Figure C.7: A subset of the categories and codes from Study 3.

C.4 Examples of coding in Study 4

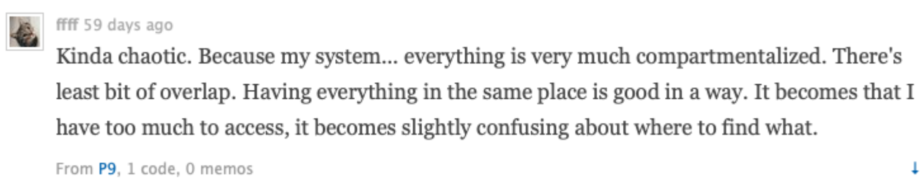
In Study 4, we first used Delve[3] for collaborative open coding and then Saturate for re-coding data with the main themes. Below, we include some example quotes focused on centralization (Figure C.8, Figure C.9) and customization (Figure C.10, Figure C.11).


Janet Chen 60 days ago

[least relevant scenario] Data leak. Tend to store less sensitive information on the cloud. It is what it is from cloud storage. It's smart that it would suggest you to encrypt that. Personally, I would stay away from storing it in a cloud platform anyway. I think it would be useful. But it depends on the person's preference and when they start learning about organizing. For me, I start with local drives. Possibly the newer generation, introduced to cloud storage off the bad. For me, it would be always local. The safest. I'm not... some people like to collaborate in one place, one app, access to email accounts and whatever, but that's not my preferred way.

From P8, 3 codes, 0 memos

Figure C.8: Example of coding about centralization from Study 4.


ffff 59 days ago

Kinda chaotic. Because my system... everything is very much compartmentalized. There's least bit of overlap. Having everything in the same place is good in a way. It becomes that I have too much to access, it becomes slightly confusing about where to find what.

From P9, 1 code, 0 memos

Figure C.9: An additional example of coding about centralization from Study 4.

[3] https://delvetool.com

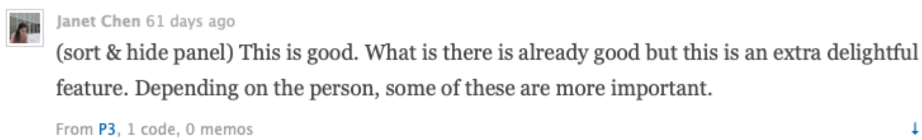


Figure C.10: Example of coding about customization from Study 4.

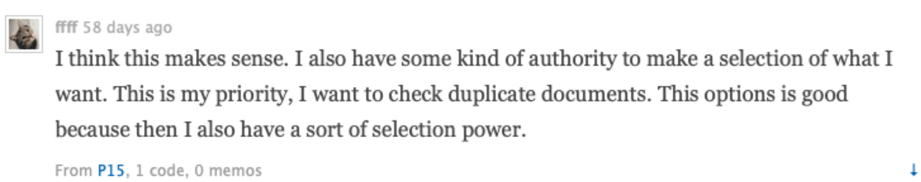


Figure C.11: An additional example of coding about customization from Study 4.

C.5 Examples of coding in the survey responses

To code the survey responses, I used Annotations. Below, some examples of coding for data types (Figure C.12) and decluttering criteria (Figure C.13).

deleting photos

Paragraph 3, Q8-final
pictures

Paragraph 4, Q8-final
Deleted photos

Paragraph 5, Q8-final
deleted some old photos

Paragraph 6, Q8-final
deleting image

Paragraph 7, Q8-final
some photos

Paragraph 9, Q8-final
pictures

Paragraph 11, Q8-final
I deleted a lot of unneeded photos

Paragraph 14, Q8-final
my girlfriends pictures with me

Paragraph 15, Q8-final
I delete pictures on my phone

Paragraph 20, Q8-final
photos

Paragraph 21, Q8-final
Pictures

Paragraph 26, Q8-final
I deleted NSFW photos that my friends send on group chats

Figure C.12: Example of coding data types in the online survey.

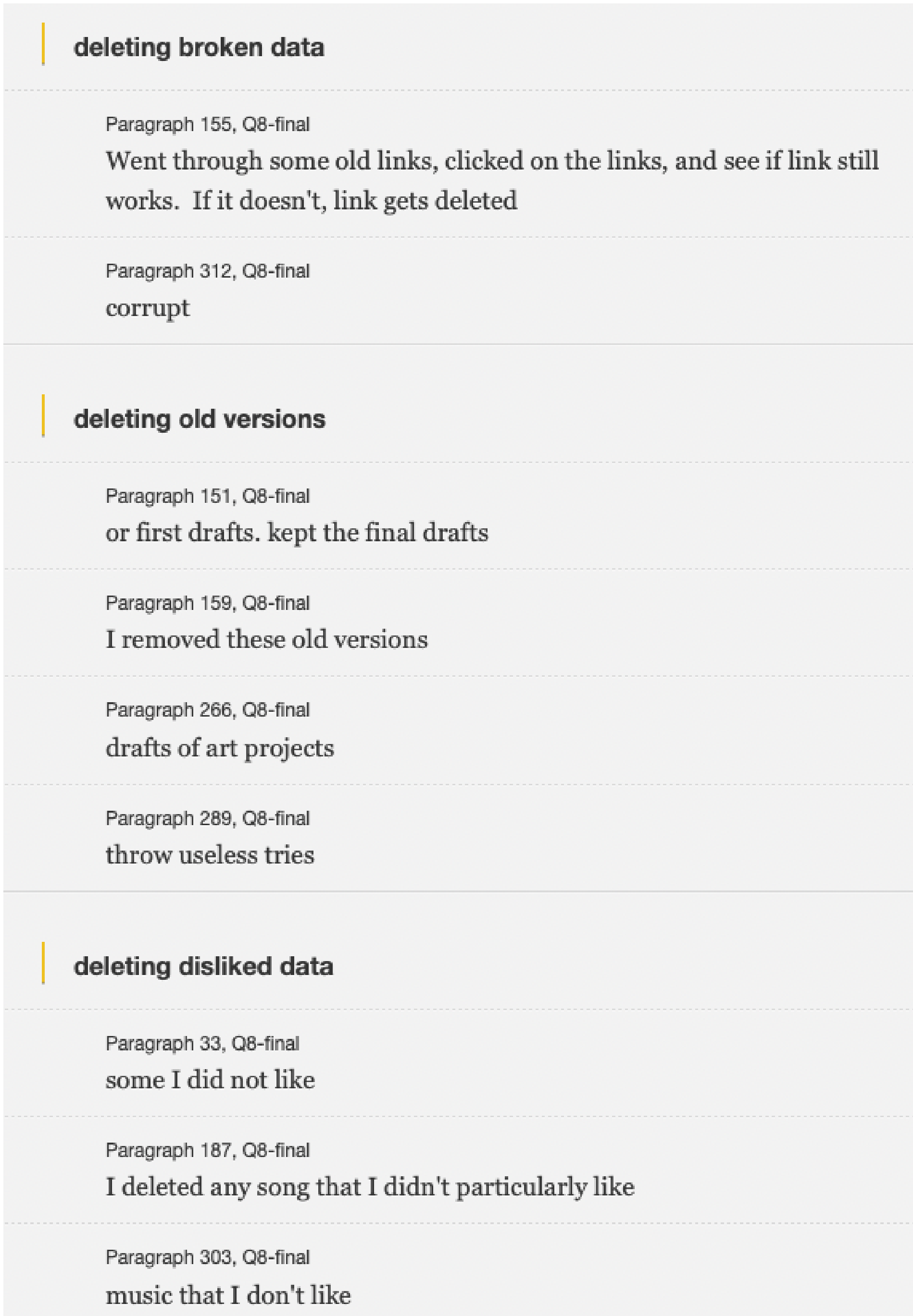


Figure C.13: Example of coding decluttering criteria in the online survey.

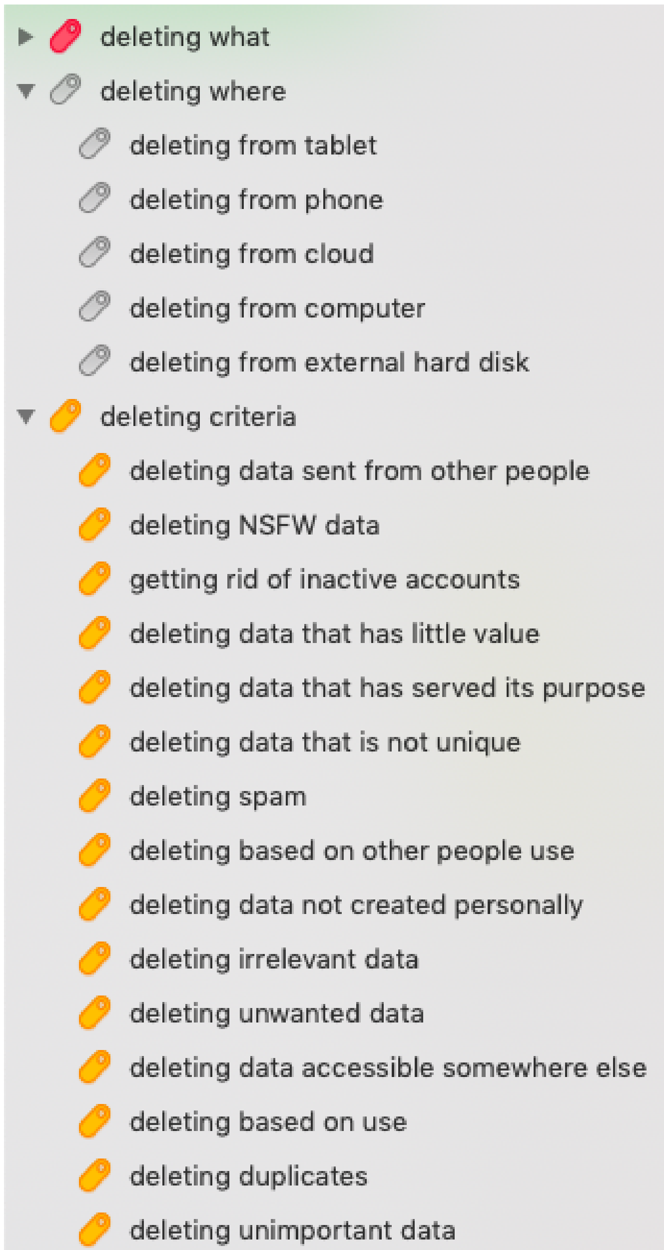


Figure C.14: A subset of the categories and codes from the survey.

Appendix D

Additional Study Materials

D.1 Screening surveys for Study 3 and Study 4

Study 3 screening survey

Start of Block: Participant information

Q1 **Data Management Tools study recruiting**

We are recruiting for an interview study on digital data management tools. Please complete the survey and we will get back to you in case you are selected, thanks!

There are no right or wrong answers in this screening survey: we are looking for a varied set of participants.

*In the interview, we will show you a series of systems to manage data in the form of short videos. Then, we will ask you questions about them. We will also ask you to bring some of your main interactive devices with you and ask about your general data management behaviors. We will audio-record the session. The study will last around **1 hour**. The compensation is **$15**.*

(For any questions, please contact REDACTED)

Q2 How old are you?

____

Q3 What is your main occupation?

____

Q4 How would you characterise your general expertise with interactive technology?

- ○ Below average
- ○ Average
- ○ Above average

Q5 Do you have experience with programming or a background in Computer Science?

- ○ Yes
- ○ No

Q6 What interactive systems do you use most frequently?

- ☐ Windows
- ☐ macOS (Apple computers)
- ☐ Android (Samsung, LG, Motorola mobile devices)
- ☐ iOS (Apple mobile devices)
- ☐ Other ____

Q7 What cloud platforms do you use most frequently?

- [ ] Dropbox
- [ ] Google Drive
- [ ] iCloud
- [ ] Box
- [ ] OneDrive
- [ ] Amazon Drive
- [ ] Other ____
- [ ] I don't use any cloud platforms

Q8
Thinking of your general approach with managing digital data (files, texts, photos, videos, etc.) what is the option that best represents you?
(*It doesn't need to be a perfect match, you can choose the one that is closest to you even if you have some exceptions*.)

◯ I keep some data, delete some, and generally have a pretty relaxed approach to organization. I don't think or worry much about data management.

◯ I keep most of my data, but sometimes feel overwhelmed by it. I am not as organized as I probably should be because it is hard or I don't have time to do it. I would like to get rid of some data.

◯ I keep lots of data, I am generally organized, and I am happy about my approach.

◯ I only keep necessary data. I am organized and regularly spend time deleting and managing data.

◯ I do not have a lot of data to begin with and I try to avoid spending a lot of time with technology.

Q9 What is your email address?
(We will use it to contact you in case you are recruited for the study)

◯ Email Address ____

End of Block: Participant information

Study 4 screening survey

Start of Block: Participant information

Q1 Data Management Tool study recruiting

We are recruiting for a study on digital data management tools. **Please complete the survey and we will get back to you in case you are selected, thanks**!
There are no right or wrong answers in this screening survey: we are looking for a varied set of participants.

In the interview, we will show you and ask you to interact with a prototype in a set of use scenarios. We will ask you questions about your experience of interacting with the prototype. We will also ask you questions on your behaviors and attitudes towards data management. We will audio-record the session. The study will last around **90 minutes**. The compensation is **$20**.

(For any questions, please contact REDACTED)

Q2 How old are you?

____

Q3 What is your main occupation?

____

Q4 How would you characterise your general expertise with interactive technology?

- ◯ Below average (1)
- ◯ Average (2)
- ◯ Above average (3)

Q5 Do you have experience with programming or a background in Computer Science? (If yes, please elaborate)

○ No (1)

○ Yes (2) ____

Q6 What interactive systems do you use most frequently?

▢ Windows (1)

▢ macOS (Apple computers) (2)

▢ Android (Samsung, LG, Motorola mobile devices) (3)

▢ iOS (Apple mobile devices) (4)

▢ Other (5) ____

Q7 What cloud platforms do you use most frequently?

- ▢ Dropbox (1)
- ▢ Google Drive (2)
- ▢ iCloud (3)
- ▢ Box (4)
- ▢ OneDrive (5)
- ▢ Amazon Drive (6)
- ▢ Other (7) ____
- ▢ I don't use any cloud platforms (8)

Q8 Have you ever used any of these tools or applications?

- ▢ CCleaner (1)
- ▢ Clean My Mac (2)
- ▢ DaisyDisk (3)
- ▢ Files by Google app on Android (4)
- ▢ Google Dashboard (5)
- ▢ Settings panels to clean up data or free up space on your computer or smartphone (6)

Q9
Thinking of your general approach with managing digital data (files, texts, photos, videos, etc.) what is the option that best represents you?
(*It doesn't need to be a perfect match, you can choose the one that is closest to you even if you have some exceptions*.)

○ I keep some data, delete some, and generally have a pretty relaxed approach to organization. I don't think or worry much about data management. (1)

○ I keep most of my data, but sometimes feel overwhelmed by it. I am not as organized as I probably should be because it is hard or I don't have time to do it. I would like to get rid of some data. (2)

○ I keep lots of data, I am generally organized, and I am happy about my approach. (3)

○ I only keep necessary data. I am organized and regularly spend time deleting and managing data. (4)

○ I do not have a lot of data to begin with and I try to avoid spending a lot of time with technology. (5)

Q10 What is your email address?
(We will use it to contact you in case you are recruited for the study)

○ Email Address (1) ____

End of Block: Participant information

Appendix E

Additional Documents

E.1 Study 1 call for participation

Advertisement everyday users

Subject: Digital Data Management Study

We are conducting a study to investigate of how people manage and preserve digital data (e.g., files on computers, data produced with mobile applications) across their personal devices. The study will consist of an interview where you will discuss how you manage your digital data, showing us examples from your devices. We will ask you to explain your data management strategies, how you categorise digital data, and how you deal with the possibility of losing digital data. We might also ask you to collect screenshots of how you organize your files or data, but you will have the freedom to choose what we can capture or not.

The interview will take approximately one hour. You will receive $15 for participating.

The study is available to anyone who meets the following:

- Is 18 years of age or older
- Regularly uses one or more interactive devices (e.g. smartphone, laptop, tablet) with personal files or data on them
- Speaks English fluently

When
Participation times are negotiable, and will take place at the UBC Computer Science building or in a convenient location of your choice (your office, a café, etc.).
If you are interested, please contact **Francesco Vitale** [redacted]**.** The principal investigator for this study is Dr. Joanna McGrenere, Professor, Computer Science, UBC**.**

E.2 Study 1 recruiting poster

THE UNIVERSITY OF BRITISH COLUMBIA

TELL US ABOUT YOUR DIGITAL DATA

(AND GET 15$)

WHAT IS IT?

We are studying how people manage and preserve digital data (e.g., files on computers, data from mobile applications) on their devices.

WHAT WILL YOU DO?

You will take part in an interview (around **1 hour**) about how you manage your digital data, showing us examples from your devices. You will receive **15$** for participating.

WHO DO WE WANT?

We are looking for participants who

- are 18 years or older
- speak English fluently
- regularly use interactive devices (e.g., smartphone, laptop, tablet)

WHEN?

Participation times are negotiable, at the UBC Computer Science building or in a location of your choice (your office, a café, etc.)

HOW TO PARTICIPATE

If interested, please contact **Francesco Vitale:** [redacted] mentioning your age, occupation, and what devices you use.

E.3 Study 1 consent form

UBC Department of Computer Science
ICICS/CS Building
201-2366 Main Mall
Vancouver, B.C., V6T 1Z4

Digital Data Management Study

Consent Form

Principal Investigator
Joanna McGrenere, Professor, Department of Computer Science, [redacted]

Co-Investigator
Francesco Vitale, PhD student, Department of Computer Science, [redacted]

Project Purpose and Procedures
The purpose of this study is to investigate how people manage, categorise and preserve their digital data (e.g., files on computers, data produced with mobile applications). During the study, we will ask you questions about your data management practices and your experience with digital data loss and preservation. You will be free to discuss details of your data in as much detail as you wish, whatever you are comfortable with. We might also ask you to collect screenshots or pictures from your devices of how you organize your data, but we will ask permission to do so and you will be free to refuse, without impacting your participation in the study.

Confidentiality
Your identity will be kept confidential. We will record the audio of your answers, take notes on your comments, and in some cases collect screenshots or pictures, but no identifying information will be stored with this data, nor will it be associated with the data after it has been analyzed.

The results will be made public through scholarly publications, presentations and academic theses; however, no identifying information will be included in any of these.

Risks/Remuneration/Compensation
We do not anticipate any significant risks for taking part in this study. You are free to withdraw at any point in your participation, even after it has been completed.

You will receive an honorarium of $15 for your participation. You will be eligible for the honorarium even if you withdraw from the study.

Contact Information about the Project
If you have any questions or require further information about the project you may contact Francesco Vitale ([redacted]), or Dr. Joanna McGrenere ([redacted]).

Contact for Concerns About the Rights of Research Subjects
If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics at [redacted] or if long distance e-mail [redacted] or call toll free [redacted].

Consent
We intend for your participation in this project to be pleasant and stress-free. Taking part in this study is entirely up to you. You have the right to refuse to participate in this study. If you decide to take part, you may choose to pull out of the study at any time without giving a reason and without any negative consequence.

Your signature below indicates that you have received a copy of this consent form for your own records.

Your signature indicates that you consent to participate in this study.

I, (print name) ___________________________ agree to participate in the study as outlined above. My participation in this study is voluntary and I understand that I may withdraw at any time.

____
Participant Signature   Date

____
Printed Name of the Participant signing above

E.4 Study 2 call for participation

Advertisement group 4 (everyday minimalist users)

Subject: Study on Minimalism

We are conducting a study to investigate how people minimize (i.e., cleanup or de-clutter) their possessions, both physical (e.g., clothes, books, tools) and digital (e.g., files on computers, data produced with mobile applications). The study will consist of an observation and interview where you will discuss how you follow a minimalist approach in managing your possessions. We will ask you to show us examples from your devices and your home or office and describe the process you go through when minimizing your possessions. We will ask you to discuss your approach in deciding what to keep and discard. We will audio and video record the session and might also ask you to collect screenshots or pictures of your items or during the minimizing process, but you will have the freedom to choose what we can capture or not.

The study will last approximately one hour. You will receive $45 for participating.

The study is available to anyone who meets the following:

- Is 18 years of age or older
- Regularly uses one or more interactive devices (e.g. smartphone, laptop, tablet) with personal files or data on them
- Speaks English

When and where
Participation times are negotiable, and will ideally take place at your home or office or another equivalent place where you store your own possessions.

If you are interested, please contact **Francesco Vitale.**The principal investigator for this study is Dr. Joanna McGrenere, Professor, Computer Science, UBC**.**

E.5 Study 2 recruiting poster

We note that the recruiting poster for Study 2 had a typo, indicating a compensation of \$45 instead of \$40.


ARE YOU A MINIMALIST?
(EARN 45$)
We are studying how self-identified "minimalists" manage their possessions. You will take part in an observation and interview (around 45 minutes, in your home) about your possessions (both physical and digital), showing us examples of how you minimize them. You will receive 45$.
WHO DO WE WANT?
We are looking for participants who
◦ are 18 years or older and speak English
◦ regularly use digital devices (e.g., smartphone, laptop)
◦ self-identify as minimalists or interested in minimalism
WHEN AND WHERE?
Participation times are negotiable. The study will take place at your home, office, or equivalent place where you store your possessions.
HOW TO PARTICIPATE
If interested, please contact Francesco Vitale: [redacted]
THE UNIVERSITY OF BRITISH COLUMBIA

E.6 Study 2 consent form

UBC Department of Computer Science
ICICS/CS Building
201-2366 Main Mall
Vancouver, B.C., V6T 1Z4

Digital Data Management Study

Consent Form

Principal Investigator
Joanna McGrenere, Professor, Department of Computer Science, [redacted]

Co-Investigator
Francesco Vitale, PhD student, Department of Computer Science, [redacted]

Project Purpose and Procedures
The purpose of this study is to investigate how people minimize (i.e., cleanup or de-clutter) their possessions, both physical (e.g., clothes, books, tools) and digital (e.g., files on computers, data produced with mobile applications). During the interview we will ask you questions about how you manage your possessions. We will also ask you to observe the process you go through to keep your possessions at a minimum. You will be free to discuss details of your possessions in as much detail as you are comfortable with.

Confidentiality
Your identity will be kept confidential. We will audio and video record the session, take notes on your comments, and in some cases collect screenshots or pictures, but no identifying information will be stored with this data (we will avoid capturing your face in video), nor will it be associated with the data after it has been analyzed.

The results will be made public through scholarly publications, presentations and academic theses; however, no identifying information will be included in any of these.

Risks/Remuneration/Compensation
We do not anticipate any significant risks for taking part in this study. You are free to withdraw at any point in your participation, even after it has been completed.

You will receive an honorarium of $40 for your participation. You will be eligible for the honorarium even if you withdraw from the study.

Contact Information about the Project
If you have any questions or require further information about the project you may contact Francesco Vitale ([redacted]), or Dr. Joanna McGrenere ([redacted]).

Contact for Concerns About the Rights of Research Subjects
If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics at [redacted] or if long distance e-mail [redacted] or call toll free [redacted].

Consent
We intend for your participation in this project to be pleasant and stress-free. Taking part in this study is entirely up to you. You have the right to refuse to participate in this study. If you decide to take part, you may choose to pull out of the study at any time without giving a reason and without any negative consequence.

Your signature below indicates that you have received a copy of this consent form for your own records.

Your signature indicates that you consent to participate in this study.

I, (print name) ____
agree to participate in the study as outlined above. My participation in this study is voluntary and I understand that I may withdraw at any time.

____
Participant Signature   Date

E.7 Study 3 call for participation

Advertisement Phase III

Subject: Help us design new data management tools

We are studying how people manage and select their digital data over time. In this interview, we will show you a series of possible systems to manage data in the form of short videos. Then, we will ask you questions about them. We will also ask you about your general data management attitudes and behaviors. We will audio-record the session.

The study will last approximately one hour. You will receive $15 for participating.

The study is available to anyone who meets the following:

- Speaks English
- Is 18 years old or older
- Regularly uses one or more interactive devices (e.g. smartphone, laptop, tablet) with personal files or data on them

When and where
Participation times are negotiable. The interview will take place at the Department of Computer Science at UBC, Vancouver campus.

If you are interested, please contact **Francesco Vitale.** The principal investigator for this study is Dr. Joanna McGrenere, Professor, Computer Science, UBC**.**

E.8 Study 3 consent form

UBC Department of Computer Science,
ICICS/CS Building, 201-2366 Main Mall
Vancouver, B.C., V6T 1Z4

Digital Data Management Study - Consent Form

Principal Investigator
Joanna McGrenere, Professor, Department of Computer Science, [redacted]

Co-Investigator
Francesco Vitale, PhD student, Department of Computer Science, [redacted]

Project Purpose and Procedures
The goal of this study is to gather feedback on possible systems to manage digital data. We will show you a series of video prototypes and ask you questions about them. We will also ask you questions on your behaviors and attitudes towards data management. You are free to answer in as much detail as you are comfortable with.

Confidentiality
Your identity will be kept confidential. We will audio record the session, take notes, and in some cases collect pictures or video, but no identifying information will be stored with this data (we will avoid capturing your face in case of pictures or video capture), nor will it be associated with the data after it has been analyzed. The results will be made public through scholarly publications, presentations and academic theses; however, no identifying information will be included in any of these.

Risks/Compensation
We do not anticipate any significant risks for taking part in the study. You are free to withdraw at any point, even after it has been completed. You will receive an honorarium of $15 for your participation. You will be eligible for the honorarium even if you withdraw from the study.

Contact Information about the Project
If you have any questions or require further information about the project you may contact Francesco Vitale ([redacted]), or Dr. Joanna McGrenere ([redacted]).

Contact for Concerns About the Rights of Research Subjects
If you have any concerns or complaints about your rights as a research participant and/or your experience while participating in the study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics at [redacted] or if long distance e-mail [redacted] or call toll free [redacted] .

Consent
We intend for your participation to be pleasant and stress-free. Taking part in this study is entirely up to you. You have the right to refuse to participate. If you decide to take part, you may choose to pull out of the study at any time without giving a reason and without any negative consequence.

Your signature below indicates that you have received a copy of this consent form for your own records. Your signature indicates that you consent to participate in this study.

I, (print name) ____

agree to participate in the study as outlined above. My participation in this study is voluntary and I understand that I may withdraw at any time.

____ ____________________

Participant Signature Date

Consent Form Phase III – Version 1.1 – August 22, 2018

E.9 Study 4 call for participation

Advertisement Phase IV

Subject: Help us evaluate a prototype to manage your personal digital data

We are studying how people manage their personal digital data (for example, files, documents, photos, messages). In this study, we will ask you to interact with the prototype of a system for managing personal data. Then, we will ask you questions about your experience with the prototype. We will also ask you about your general data management attitudes and behaviors. We will audio-record the session.

The study will last approximately 60-90 minutes. You will receive $20 for participating.

The study is available to anyone who meets the following:

- Speaks English
- Is 18 years old or older
- Regularly uses one or more interactive devices (e.g. smartphone, laptop, tablet) with personal files or data on them

When and where
Participation times are negotiable. The interview will take place at the Department of Computer Science at UBC, Vancouver campus.

If you are interested, please contact **Francesco Vitale.** The principal investigator for this study is Dr. Joanna McGrenere, Professor, Computer Science, UBC**.**

E.10 Study 4 consent form

UBC Department of Computer Science,
ICICS/CS Building, 201-2366 Main Mall
Vancouver, B.C., V6T 1Z4

Digital Data Management Study - Consent Form

Principal Investigator
Joanna McGrenere, Professor, Department of Computer Science, [redacted]

Co-Investigator
Francesco Vitale, PhD student, Department of Computer Science, [redacted]

Project Purpose and Procedures
The goal of this study is to gather feedback on a system to manage digital data. We will show you and ask you to interact with a prototype in a set of use scenarios. We will ask you questions about your experience of interacting with the prototype. We will also ask you questions on your behaviors and attitudes towards data management. You are free to answer in as much detail as you are comfortable with.

Confidentiality
Your identity will be kept confidential. We will audio record the session, take notes, and in some cases collect pictures or video, but no identifying information will be stored with this data (we will avoid capturing your face in case of pictures or video capture), nor will it be associated with the data after it has been analyzed. The results will be made public through scholarly publications, presentations and academic theses; however, no identifying information will be included in any of these.

Risks/Compensation
We do not anticipate any significant risks for taking part in the study. You are free to withdraw at any point, even after it has been completed. You will receive an honorarium of $20 for your participation. You will be eligible for the honorarium even if you withdraw from the study.

Contact Information about the Project
If you have any questions or require further information about the project you may contact Francesco Vitale ([redacted]), or Dr. Joanna McGrenere ([redacted]).

Contact for Concerns About the Rights of Research Subjects
If you have any concerns or complaints about your rights as a research participant and/or your experience while participating in the study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics at [redacted] or if long distance e-mail [redacted] or call toll free [redacted] .

Consent
We intend for your participation to be pleasant and stress-free. Taking part in this study is entirely up to you. You have the right to refuse to participate. If you decide to take part, you may choose to pull out of the study at any time without giving a reason and without any negative consequence.

Your signature below indicates that you have received a copy of this consent form for your own records. Your signature indicates that you consent to participate in this study.

I, (print name) ____

agree to participate in the study as outlined above. My participation in this study is voluntary and I understand that I may withdraw at any time.

____ ____________________

Participant Signature Date

Ethics application H17-00734 - Consent Form Phase IV – Version 1.2 – August 6, 2019

E.11 Survey call for participation

Survey on Digital Data
Researcher: Francesco Vitale (graduate student), Joanna McGrenere (PI)
Description: We are studying how people manage and preserve digital data (e.g., files on computers, photos, apps.) In the survey we will ask you questions about your digital data preservation startegies and organization. The survey should take 10-15 minutes to complete and you can find it at this address: https://ubc.ca1.qualtrics.com/jfe/form/SV_3fW4jIZYUVnO6UZ
Eligibility: We are looking for participants who are 18+ years old.
Location: Take the survey on a device of your choice, wherever you are.
Contact Information: Please contact Francesco Vitale, using "Digital Data Survey" as a reference if you have any questions
Reimbursement/Time: You can enter a prize draw for a $25 (CAD) gift card. We will extract one person out of every 10 participants.
Study End Date: December 30, 2018

E.12 Survey consent section

UBC Department of Computer Science
ICICS/CS Building
201-2366 Main Mall
Vancouver, B.C., V6T 1Z4

Digital Data Management Study

Consent Form Survey

Principal Investigator
Joanna McGrenere, Professor, Department of Computer Science, [redacted]

Co-Investigator
Francesco Vitale, PhD student, Department of Computer Science, [redacted]

Project Purpose and Procedures
The purpose of this study is to investigate how people manage, categorise and preserve their digital data (e.g., files on computers, data produced with mobile applications). During the survey, we will ask you questions about your digital data.

Confidentiality
Your identity will be kept confidential. We will only record your identifying information if you want to participate in the draw for a prize. Your identifying information will not be stored with this data, nor will it be associated with the data after it has been analyzed.

The results will be made public through scholarly publications, presentations and academic theses; however, no identifying information will be included in any of these.

Risks/Remuneration/Compensation
We do not anticipate any significant risks for taking part in this study. You are free to withdraw at any point in your participation. Participants will be able to enter a draw with a 1/10 chance of winning a $25 (25 Canadian dollars) gift card. Everyone who participates (even those who withdraw) is allowed to take part in the draw for a prize.

Contact Information about the Project
If you have any questions or require further information about the project you may contact Francesco Vitale ([redacted]), or Dr. Joanna McGrenere ([redacted]).

Contact for Concerns About the Rights of Research Subjects
If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics at [redacted] or if long distance e-mail [redacted] or call toll free [redacted].

Consent
We intend for your participation in this project to be pleasant and stress-free. Taking part in this study is entirely up to you. You have the right to refuse to participate in this study. If you decide to take part, you may choose to pull out of the study at any time without giving a reason and without any negative consequence.

By answering the questions in the survey you agree to take part in the study.