FACILITATING DATA SHARING:
A DESIGN APPROACH TO INCORPORATE CONTEXT INTO THE RESEARCH DATA REPOSITORY

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Engineering and Physical Sciences

2016

By
Kristian Garza
School of Computer Science
## Contents

1. **Abstract** 12
2. **Declaration** 13
3. **Copyright** 14
4. **Acknowledgements** 15

5. **1 Introduction** 16
   1.1 **Research Hypotheses** 18
   1.2 **Research Aims and Objectives** 19
   1.3 **Methodology and Approach** 20
   1.4 **Research Contributions** 20
   1.5 **Publications and Research Activity** 23
      1.5.1 **Published Work** 23
      1.5.2 **Research Activity** 23
   1.6 **Thesis Organisation** 23

6. **2 Background** 25
   2.1 **Scientific Data Sharing** 26
   2.2 **Small Scientific Collaborations** 29
   2.3 **Science Data Repositories** 30
   2.4 **Challenges for Small Collaborations** 33
   2.5 **Coping with Data Withholding** 34
   2.6 **Related Work** 38
      2.6.1 **SDR Development Life Cycle** 39
      2.6.2 **Context Capturing** 40
      2.6.3 **Adapting Strategy** 42
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Differences of data-sharing definitions from the literature.</td>
<td>27</td>
</tr>
<tr>
<td>2.2</td>
<td>Different Mediums to Share Data and their features.</td>
<td>31</td>
</tr>
<tr>
<td>2.3</td>
<td>Examples of Scientific Data Repositories</td>
<td>32</td>
</tr>
<tr>
<td>2.4</td>
<td>Comparison of approaches to capture and incorporate context into SDRs.</td>
<td>38</td>
</tr>
<tr>
<td>3.1</td>
<td>Embedded cases descriptive information. Here we compare and contrast the six collaboration that were studied in this work.</td>
<td>56</td>
</tr>
<tr>
<td>3.2</td>
<td>Interview guide items. We categorise the interview guide’s item by each construct item.</td>
<td>57</td>
</tr>
<tr>
<td>3.3</td>
<td>Interviewees demographics (n=10). The study’s participants are a balanced representation of the demographics of the collaborations.</td>
<td>58</td>
</tr>
<tr>
<td>3.4</td>
<td>Comparison of reported collaboration’s contributions as of 2015.</td>
<td>62</td>
</tr>
<tr>
<td>3.5</td>
<td>Comparison between last publication and last activity in the SDR.</td>
<td>69</td>
</tr>
<tr>
<td>3.6</td>
<td>Exploratory Study Findings.</td>
<td>74</td>
</tr>
<tr>
<td>3.7</td>
<td>Categorisation of factors associated with the data-sharing in the literature.</td>
<td>88</td>
</tr>
<tr>
<td>4.1</td>
<td>Survey of information systems contextualised in the Choice-Architecture framework from Munscher et al. Illustrating the processes and mechanism of Choice-Architecture in the wild.</td>
<td>97</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison of elements reframed in the experimental set-up of the SDR. The reframing was performed using Emphais Framing.</td>
<td>112</td>
</tr>
<tr>
<td>4.3</td>
<td>Experimental Sequences Comparison and Participants demographics. Showing a balance between sequences due to the pseudo-randomisation.</td>
<td>133</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison of Accessibility Results. Using three different CLM Models. Showing the effect of the Emphasis Frame intervention.</td>
<td>134</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison of the Usability Results. Using the Cumulative Link Mixed Model.</td>
<td>137</td>
</tr>
</tbody>
</table>
List of Figures

1.1 A readers’ guide to the structure, dependencies and deliverables of this thesis. Each chapter represents the application and evaluation of the approaches generated by this work. 21

2.1 SDR development lifecycle. Highlight the stages for context incorporation. 39

2.2 This thesis’s socio-technical method using an analytic approach to context-capturing. This approach complements the empirical strategy used in the literature. 45

2.3 This thesis’s persuasive design approach for context-incorporation in an SDR. 48

2.4 Organisation of the application of the Persuasive Design Approach in terms of the chapters ahead. 50

3.1 Empirical and Analytic approaches to context-capturing. We present the three studies presented in this chapter in the context of the approach they serve. 53

3.2 Interviews participation. Showing a 17% drop-off rate. 59

3.3 Data Published and data shared with collaborators (i.e., Specific Sharing) from the six collaborations. Most of the collaboration kept their data hidden away from the public and published less than 25%. 68

3.4 Distribution of self-depositions over time by the six collaborations. Showing a relationship between rapid increases of self-deposition and funding evaluation deadlines. 71

3.5 Decision tree on the alternative studies considered to the exploratory case study. The study conducted in subsection 3.1.2 is highlighted in green. The studies piloted are highlighted in yellow. 81
3.6 Decision tree on the alternative studies considered to Measuring Data Sharing. The study conducted in subsection 3.1.5 is highlighted in green. The studies piloted are highlighted in yellow.

4.1 In this chapter we focus on the design stage of the SDR development life cycle.

4.2 Github's dashboard making user’s own behaviour visible with a calendar heatmap.

4.3 Academia.edu landing page after signup references a publication from PLOS ONE that associates manuscripts depositing in Academia.edu with citations boost. This way Academia.org nudges subjects into be cataloging publications on the site.

4.4 Figshare (left) and Zenodo (right) data repositories services are remarkable similar but with a different spin to data-sharing. Figshare focused on obtaining credit while Zenodo is more neutral.

4.5 ResearchGate’s user feed with the latest publication registration from your collaborators. In this way ResearchGate establishes the descriptive norm: “Keep posting.”

4.6 F1000Research landing page referring to leaders’ opinions on the subject of Open Access.

4.7 Zenodo.org employs a no-action default for open access in its data upload form. In this way the form nudges user to select the open access option.

4.8 The sharing preferences setting in the upload form of the SEEK4Science system. The grouping of the current options offer great flexibility (18 combinations) but a regrouping of option could improve the users decision making.

4.9 Figshare.com upload page with DOI pre-assignation on upload. DOI pre-assignation is a benefit users can get when depositing data in Figshare.

4.10 Academia.edu reminding users to visit their service.

4.11 Context-incorporation using the Choice-Architecture based procedure. We include the two empirical claims that will be used in the next section to define design features.
4.12 Contextualising the Context-Incorporation Stage. This stage transforms the findings identified in the context capturing stage into design claims for the SDR (i.e., SEEK).

4.13 SEEK4Science landing page completely reframed to highlight data-sharing as a task that could lead to obtaining citations.


4.15 Design Claim DC2 represent the implementation of a combinations of techniques: establish grouping and refering to a social norm.

4.16 Experimental protocol as object of study of this pilot. Abbreviations: Pt.: Participant; Random.:Randomization.

4.17 Full Crossover Trial Protocol. In the crossover protocol the measurements are repeated but the second time the treatments are administrated in reverse order. Abbreviations: Pt: Participant; Random: Randomization.

4.18 Participants breakdown. Showing a 20% response rate.

4.19 The effect on the accessibility means over the two periods seem to be positive. Showing a non statistical significant carryover effect.

4.20 Effect on the usability means over the two periods. Showing a non statistical significant carryover effect.

4.21 Sequence group effect on the regret means over the two periods. Showing a border line significant carryover effect.

4.22 Data accessibility preference comparison under two conditions. Control condition (top) and Treatment condition (bottom). The participants under the Treatment condition \((A)\) selected more open data accessibility options for the data they uploaded. The response is given in terms of the accessibility setting selected.

4.23 Aggregated participants embargo Periods preference comparison under two conditions. Control condition (top) and treatment condition (bottom). Participants under the Treatment condition \((A)\) selected shorter embargo periods for the data they uploaded in the experiment. Response is given in terms of the length of the embargo period selected.
4.2.4 Decision tree on the alternative studies considered to the experimental crossover study. The study conducted in subsection 4.5.2 is highlighted in green. The studies piloted are highlighted in yellow.

5.1 Structure of the thesis presented

C.1 Poster and flyer invitation used to recruit participants

C.2 Control condition’s landing page. All the textual information was framed to fit the current frame of the SDR.

C.3 Control condition’s upload form. Two intervention were added in this design, an Emphasis Frame for the accessibility options and we avoided to use grouping and classify for the embargo period selection.
Abstract

We asked whether the design of an Science Data Repository (SDR) can influence data sharing behaviour in small scientific collaborations. We hypothesised that an SDR can influence data-sharing behaviour when its design considers the context of data-sharing. We proposed an alternative approach to those documented in the literature, employing a combination of socio-technical empirical and analytical methods for context capturing, and choice architecture for context incorporation. To evaluate the approach we applied it to design features in a Scientific Data Repository for a population of small scientific collaborations within the Life Sciences.

The application of this thesis’ approach consisted of an exploratory case study, a review of factors associated with data sharing, the definition of design claims, and implementation of a set of design features. We collected data using interviews with members of the collaborations and designers of the SDR; as well as obtaining the data-logs from the collaborations’ SDR. We evaluated the resulting design features using an asynchronous web experiment.

We found that using the empirical approach to context capturing we are able to effectively identify factors associated with data sharing in the small scientific collaborations. Moreover, we identified a number of limitations on the application of the analytical approach to context capturing. Furthermore, we found that the Choice Architecture based procedure for context incorporation can define effective design features in Science Data Repositories.

In this work, we show that we can facilitate data-sharing by incorporating context into the design of a Science Data Repository, and identified a set of restrictions to use our approach. The approach proposed in this thesis can be used by practitioners wishing to improve data sharing in an SDR. Contributions, such as the survey of factors associated with data sharing behaviour, can be used by researchers to understand the problems associated with data sharing in small scientific collaborations.
Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
First and foremost, I would like to thank my parents Teresa and Javier; whose support has been truly invaluable. Thank you for enabling me to grow as a person. I cannot say thank you enough for being so supportive and caring throughout all these years that I have been away.

Five years ago, I sent a small research proposal with the wish that would arise the interest of some people. It did. Those people were Prof. Carole Goble and Dr. John Brooke. Carole, you allowed me a great deal of independence and backed it up with unfailing encouragement, and support. You have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a researcher. John, your advice has been priceless and I would like to thank you for your patience, motivation, and sharing immense knowledge with me.

Thanks to Dr. Caroline Jay; whose thought-provoking discussions and guidance helped me to steer my work in the right direction. Caroline, thank you so much for joining the project your contribution was crucial.

Thanks to Rahma, without whom I’d never have started. To all my friends, especially to David, Fernando, Harriet, Indy, Kathleen, Soheir, Sofie, Takis, Tim, and Zeqian; for their help and support throughout these last couple months.

Finally, thanks to Moozhan, without whom I’d never have finished.
Chapter 1

Introduction

Scientific data are, without a doubt, pivotal components in the life cycle of knowledge production. For one thing, scientific data are of the utmost importance for the support and confirmation of scientific claims. Moreover, scientific data constitute an output in the process of scientific inquiry.

Equally important to the process of scientific inquiry is making these data openly available to others. Economically, the open provision of data yields a significant return of investment \cite{PVW11}, minimises duplication of effort \cite{Whi11}, \cite{WRB13} and increases credit between researchers \cite{Whi11}. Similarly, it furthers the progress of science by allowing the scientific community to ask new questions of extant data \cite{Bor11}, \cite{Whi11}, and it shortens the length of a cycle of the scientific process \cite{GRB13}. Moreover, it strengthens the scientific process by allowing the re-analysis of evidence \cite{Whi11}, \cite{WRB13}; the verification of results \cite{Whi11}; and the replication of research \cite{Bor11}, \cite{Roc+14}. Furthermore, open provision of scientific datasets obtained with publicly-funded money is widely considered an ethical obligation \cite{Bor11}, \cite{WRB13}.

However, the practice of open provision of datasets, or data-sharing, has not met the empirical reality that its importance would presuppose. Longitudinal studies across different fields of science have shown that data-sharing is not a common practice \cite{Ten+11a}, \cite{Ten+15}, \cite{Sta11}. Similar results have been uncovered by case studies in the Physical Sciences \cite{Cra+10}, Engineering \cite{WRB13}, and Life Sciences \cite{RIN08}, \cite{RJS11}. The same studies have brought to light that sharing mostly occurs within restricted circles \cite{Ten+11a}, \cite{Ten+15}, \cite{KH09}, \cite{Cra+10}, \cite{Vel13}, \cite{RIN08}.

Consequently, the scientific community has been addressing this misalignment
between data-sharing practice and expectations from a number of perspectives (i.e., Policies, Mandates, Incentives and Infrastructure). Funding bodies and institutions have implemented policies that established the best practices for scientists to share their data [RC08] [MP13]. Journals have established mandates that require scientists to share their data in order to publish their articles [AA+11] [SV09]. New incentives have been developed to encourage researchers to share their data [IC11] [DAD14]. Governments and institutions are building new infrastructures, the implementation cost of which are being included in funding packages.

Nevertheless, all these solutions have been encountered with resistance. Policies and mandates have had limited impact across disciplines and nations [MP13] [SV09]. Moreover, neither have new incentive mechanisms gained critical mass [MN15] nor has the infrastructure been adopted by researchers [Ten+11a] [Ten+15].

As a consequence, recommendations have been provided to improve data-sharing from all these four areas. The Computer Science community has been focused on the recommendations to improve infrastructure. These recommendations may be divided into four groups; these are: Standards [BB03], Controls [ELI+14] [Ros+12], Provenance [ELI+14] and Scientific Data Repositories (SDRs) [BB03] [Hei08] [Sto10] [WRB13] [Vel13] [ELI+14] [FFH15]. Although these four lines of investigation are all-important, Scientific Data Repositories, stand out because they are the main medium of data dissemination [Abr+13]. Moreover, in recent times, data repositories have shown themselves to be a successful tool for dissemination of digital data in diverse areas. For instance, the ENCODE Consortium provides datasets on the analysis of the entire human genome via their repository \(^1\). This data collection was constructed over more than five years, and, to date, supports 464 published experimental results in the field of Genomics [ENC04].

Accordingly, recommendations for repositories have been focused on design and may be classified in terms of non-functional characteristics [SAZ04]: (1) ease-of-use and (2) social context. A number of authors have recommended the design and implementation of easy-to-use [FFH15] and cost-effective [Hei08] data repositories with high-level exposure to all scientists [WRB13] [ELI+14]. On the other hand, authors have highlighted that the design of repositories needs to accommodate for the social context, that is: the social roles of data [BB03].

\(^1\)www.encodeproject.org
individual and field specific dynamics of data \cite{Vel13}, and the researchers concerns against data-sharing goals \cite{Sto10}. This last set of recommendations has received specific support from empirical studies on data-sharing, which suggest that, in order to improve data-sharing behaviour, data repositories need to: (1) procure attention to social factors (such as individual effort, and sharing conditions) \cite{WRB13}; (2) identify demographic and disciplinary level difference in data-sharing practices \cite{Ten+15}; and (3) address attitudes \cite{KS15}.

Despite the fact of this convergence of recommendations, there are a number of questions that still need to be answered. First, to the best of our knowledge, no study has been conducted to explore the extent to which these recommendations have trickled down to the design methodologies of Data Repositories. Second, so far as we can tell, the design practices that Data Repositories designers employ to incorporate the social context are rarely documented (e.g. \cite{Wol+11b}, \cite{Lin+08} \cite{Las10}, \cite{Ros+12}, \cite{Lag+14}). Furthermore, so far as we can tell, in the case of Data repositories design, no approaches that could streamline the incorporation of social context into the design have been proposed.

Therefore, and based on the above, it is evident that there is a pressing need to address such questions. Accordingly, in this thesis, we aim to address these questions by investigating the limitations of the current practices for Data Repository design. We also aim to define an alternative approach to incorporating the social context of data-sharing in the design of information systems so that the sharing of scientific data becomes a reality.

### 1.1 Research Hypotheses

Our central hypothesis \textbf{H1} is driven by the following research question:

\textbf{How can SDR design influence data-sharing behaviour?}

\textbf{H1}: We can use the accumulated knowledge of social psychology and studies on scientific data-sharing to devise interventions that effectively facilitate data-sharing behaviour in an SDR.

From this central hypothesis, we have developed four further research questions and their related hypotheses. The first one is:

Which type of these theories are the best suited candidates to be translated? Bearing in mind that our goal is to facilitate effective data-sharing.
1.2. RESEARCH AIMS AND OBJECTIVES

H2: There are specific aspects that relate scientific data-sharing with the Economic context of scientific research, that means we can use theories from Economics to define design claims.

What type of techniques can we effectively use to transform empirical claims into design interventions?

H3: It has been established that data-sharing is unfavourable to scientists’ individual benefit, therefore they need to be persuaded to share their data. Then we can use Choice Architecture techniques (i.e., Persuasive Design) coupled with empirical claims of effective data-sharing to implement interventions.

How can practitioners incorporate the techniques we provide into their design process?

H4: We can produce a design procedure that includes Choice Architecture techniques to design features for SDRs.

1.2 Research Aims and Objectives

The aim of this research is to investigate and develop approaches that successfully support data-sharing into the design of SDRs for small collaborations. To this end, we borrow knowledge from Social Psychology and Behavioural Economics as well as with grounding our work in empirical studies on scientific collaboration and scientific data-sharing.

Specifically, our objectives are:

• Ob1. To investigate and review the scientific community approach to data-sharing.
• Ob2. To review existing approaches to SDR design.
• Ob3. To identify specific steps in the SDR development life cycle that could be changed to improve data-sharing.
• Ob4. In light of Ob3, to design/refine an approach that, theoretically, would improve data-sharing in self-depositing SDRs.
• Ob5. To evaluate the effectiveness of the proposed approach in Ob4.
1.3 Methodology and Approach

In essence, we proposed and studied the application of a new approach to design Scientific Data Repositories. We separated the approach proposal and evaluation in three chapters. The proposal was informed by a literature review and interviews with experts from the field of scholarly communications. The first part of the evaluation was carried out using an exploratory case study, literature review and analytical case study. The second part of the evaluation was conducted using a remote web experiment. Figure 1.1 provides the reader with a guide to the structure of this thesis, beginning with this Introduction. A background and literature review allowed us to identify common features that current design methods need to address (Objectives Ob1, Ob2, Ob3). In Chapter 2 we propose a set of potential solutions to address these gaps (Objective Ob4). We tackle the evaluation of our proposed solutions with two studies, as shown in Figure 1.1.

- Study 1: Identifying and translating factors associated with data-sharing into a specific case of a Small Scientific Collaboration using an SDR.
- Study 2: Investigating SDR Interventions based on a Persuasive Design Approach.

Study 1

We conducted an exploratory case-embedded study in a set of small collaborations and review of factors associated with data sharing within the Life Sciences. We collected data using interviews with members of the collaborations and designers of the SDR, as well as obtaining the data-logs from the collaborations’ SDR.

Study 2

Using our proposed procedure to incorporating context into the SDR design, we defined a set of interventions for a Life Sciences SDR. Then we defined a controlled web experiment in order to test these interventions (Objective Ob5).

1.4 Research Contributions

The evaluation of our approach to design Scientific Data Repositories for small collaborations suggested that our approach can effectively improve scientists’
1.4. RESEARCH CONTRIBUTIONS

Figure 1.1: A readers’ guide to the structure, dependencies and deliverables of this thesis. Each chapter represents the application and evaluation of the approaches generated by this work.
data-sharing behaviour. Additionally, a number of findings have sprung out from the studies we have performed. Our findings are not only of interest to SDR designers; they also inform choice architecture researchers and researchers studying data-sharing by providing insights into opportunities and challenges in their areas of expertise.

In this section, we detail specifically the novel contributions to the state of the art that have resulted from the findings in this research.

**C1:** A Set of factors associated with data-sharing behaviour. We contributed to a survey of claims on improving data-sharing practices categorised by the level of identified empirical evidence supporting them. This survey contribution identifies the gaps for empirical evidence in factors associated with data-sharing behaviour.

**C2:** A procedure to create persuasive interventions in an SDR. We contributed a Choice-Architecture-based procedure to create design claims and interventions for Scientific Data Repositories. This is a methodological contribution for practitioners which will enable SDR designers to incorporate the context of data-sharing into its design.

**C3:** We found that Emphasis Framing can help to modify data-sharing behaviour. We contribute to the body of knowledge of Choice Architecture a previously unexplored effect of framing in scientists contributing to SDRs. This is an empirical contribution that adds to the knowledge body of framing effects and choice architecture literature for decision problems classified as social dilemmas with expert subjects.

**C4:** Identified a methodological limitation for the analytic approach of context-capturing. We contribute by assessing the application of the analytic version of socio-technical approach to small scientific collaborations using Science Data Repositories. This is a methodological contribution for practitioners that reports on the limitations of the approach in its application with small scientific collaborations.

**C5:** Identification of the use of cognitive biases in SDRs. We contribute a survey of SDRs that have non-purposely relied on

---

2We classified our contributions using the classification provided by *Research Contribution Types in Human- Computer Interaction* Jacob O. Wobbrock, Ph.D. The Information School University of Washington
the cognitive biases that choice architecture uses. This survey contribution identifies the trends on the use of subjects’ cognitive biases by SDRs designers and the gaps in the use of Choice Architecture techniques.

1.5 Publications and Research Activity

1.5.1 Published Work

In the following publications the author of this thesis was the primary contributor, or made a significant contribution to the work and text presented.


1.5.2 Research Activity


1.6 Thesis Organisation

The rest of this thesis is presented as shown in Figure 1.1. Chapter 2 presents the research background, the related work and outlines our most significant contribution: A design approach to incorporate context into the SDR. The approach is described in general terms, and we used the subsequent chapters to describe this in more detail. The following chapters present the application and evaluation of the approach divided in two stages. In Chapter 3 we present the application of the context-capturing stage in a case of a small collaboration from
the Life Sciences. Additionally, we present the first of our studies, which evaluates the context-capturing stage. In Chapter 4, we introduce the application of our procedure to design interventions and incorporating context using Choice Architecture. In addition, we present the evaluation of the design features using a remote pseudo-randomised control trial experiment. Finally in Chapter 5, we draw together conclusions from the application of our approach and the findings from the two investigations, and we propose an agenda for future work.
Chapter 2

Background

Please upload your data to the web repository

... reads the email’s title in Anna’s inbox. Anna is a wet-lab Biologist working in the UK as part of a European research collaboration. The e-mail comes from her collaborator, Hans from Germany, and the email is not solely directed to her but to all the 12 researchers that make her collaboration.

Five days later:

Hans visits the web repository for the second time and finds out that Anna is the first one to upload a dataset to the web repository. He is about to click the download button before he finds out it has been disabled.

One hour later:

Please upload your data to the web repository **without restrictions**

... reads a new email in Anna’s inbox.

This story is not unique to our case. Scientific Data-Sharing is a practice that has not permeated throughout all corners of Science and it is the focal point of a multitude of research efforts. We conducted this research under the argument that scientists would be more willing to share their data if their information systems were designed under a methodology that would promote data-sharing
by incorporating context into the design of these information systems. To drive this argument “home” we need to clarify a number of concepts and processes that involve scientists, information systems and the challenges of data-sharing. This chapter introduces these concepts and contrasts different approaches that designers have applied to address the data-sharing challenges through the design of information systems.

We argue that these design approaches do not acknowledge the lack of benefits and the perception of risks and costs associated with data-sharing in a Scientific Data Repository (SDR) – in other words the context of data-sharing. We identified four gaps in the context incorporation practices currently used in SDR designs, and we propose a modification of current methodologies and approaches. Our proposal was grounded in the SDR Development Life Cycle, and we focused on two stages of the development life-cycle: research and design. We proposed a socio-technical approach supplemented by an analytic approach for context-capturing and a “persuasive design” approach to context incorporation. Our review of data sharing in small scientific collaborations indicates that the work proposed in this thesis can address the research gaps we have identified.

This chapter begins by providing the definition of three concepts around which the work of this thesis develops: Scientific data-sharing, Small Scientific Collaborations and Scientific Data Repositories (SDRs). We then present the solutions which the scientific community has been putting forward in order to fight the data-sharing dilemma. We contrast technical solutions from mandates and incentives. After this, we outline the particular challenges and opportunities to Scientific Data Repository design posed by Common Scientific Collaborations and the scientific context of data-sharing. Lastly, we review the state of the art on SDR design approaches and the state of the art on persuasive design approaches. In our review, we identify the principal components of existing SDR design solutions and highlight practices that can inform a better SDR design approach.

2.1 Scientific Data Sharing

Data-sharing comes in numerous forms; it can describe different activities in different dimensions. In modern scientific research, data-sharing is one of many activities; researchers are expected to carry out. The term implies the action of making one’s research data available to other individuals or groups to enable
research to be reproduced, or data to be reused. Under that basic definition, there could be many permutations concerning: (1) the time at which the data are made accessible, (2) the actors with access to the data, and (3) the medium of data dissemination (or data-sharing). Table 2.1 presents four definitions of data-sharing from the literature and their differences in these three dimensions. In this section, we determine the scope under which we use the term “data-sharing”.

In this work, when we talk about data-sharing, we refer to the action of making scientific data persistently available, under unrestricted access, through an information system that is accessible via the web. As pointed out by Kratz and Strasser [KS14] in this sense, the term data-sharing overlaps with terms like Data Publication, Data Release, and Open Data. Unlike Kratz and Strasser [KS14], however, in this thesis the use of the term does not encompass the validation aspect included in their Data Publication definition.

Examples of our definition include:

- Spreadsheets with tabular data shared in the SysMO SEEK catalogue, such as Gene expression rates

- Mathematical models shared in the Virtual Liver Network repository, such as Kinetic Models of Hepatic Glucose Metabolism

- Data that support publications shared in the Zenodo service, such as strain-rate sensitivity data

In this research, the scope of the term data-sharing excludes three cases in

---


---

<table>
<thead>
<tr>
<th>Sharing Type</th>
<th>Results’ Status</th>
<th>Access</th>
<th>Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Unpublished</td>
<td>Public</td>
<td>Repository and Webpost</td>
</tr>
<tr>
<td>Specific</td>
<td>Unpublished</td>
<td>Collaborators</td>
<td>Email</td>
</tr>
<tr>
<td>Publishing</td>
<td>Un- and Published</td>
<td>Public</td>
<td>Repository</td>
</tr>
<tr>
<td>Directed</td>
<td>Un- and Published</td>
<td>Collaborators</td>
<td>Email</td>
</tr>
</tbody>
</table>

* (Haennier et al. 2013)  
* (Kratz and Strasser 2014)  
* (Piwowar 2011)
which scientists share data: Sharing upon request via e-mail, web-posting and sharing within institutional boundaries.

1. When we talk about data-sharing, we are not talking about sharing data via e-mail. Data-sharing via email has been shown to be one of the most prominent ways to share data from published articles. For instance, Cragin et al. in their case study of small science collaborations in the natural science \[\text{Cra+10}\] found that the most common way scientists shared data outside the walls of their collaboration group was by individual email requests. Similarly, in their empirical study of data-sharing willingness, Savage and Vickers \[\text{SV09}\] used email requests to obtain open data. However, data-sharing by email does not make data discoverable and is thus excluded from our definition.

2. Another case we exclude is sharing data via web-posting. When performed before the article publication, Haeussler calls web posting, “general sharing” \[\text{Hae+13}\]. In their case study of the Centre for Embedded Network Sensing, Wallis et al. \[\text{WRB13}\] found that a preferred practice among researchers was to make a dataset available locally and send the data requestor a link. As with e-mail, this practice does not enable data to be discoverable.

3. When we talk about data-sharing, we are not talking about sharing data with scientists within one’s group, project, institution or collaboration. Other terms that identify this kind of sharing are Specific Sharing \[\text{Hae+13}\] (when conducted before article publication), and Directed Sharing \[\text{Piw11}\]. This type of data-sharing is present in most research endeavour and represents a common practice within the boundaries of a research group, project, institution or collaboration. The Cragin et al. case study of different small science collaborations in the natural sciences \[\text{Cra+10}\] showed that sharing within collaborations is the major type of data-sharing those scientists conducted. Data-sharing in this scenario is restricted; hence it does not fulfil our requirements for data-sharing.
2.2 Small Scientific Collaborations

The next concept to define is Scientific Collaboration – its meaning, the different types of scientific collaborations the community has identified and their characteristics.

When we talk about the phenomenon of Scientific Collaboration, we use the definition provided by Sonnenwald \cite{Son07} as it is the one which is most related to our understanding of Scientific Collaborations. In her thorough review of scientific collaborations, Sonnenwald defines the Scientific collaboration “as the interaction taking place within a social context among two or more scientists that facilitates the sharing of meaning and completion of tasks concerning a mutually shared, superordinate goal”. Examples of collaborations under this definition include:

- The Collaboration on Diagnosis of Emerging Viruses (CoDEV)\footnote{http://www.nottingham.ac.uk/codev/codev-collaborators.aspx}: A BBSRC-funded UK-India collaborative project that aims to establish strong research partnerships between the UK and India in the field of novel diagnostic assay development for emerging viral diseases. The collaboration consists of seven contributing scientists from four universities across two countries.

- The Virtual Liver Network \cite{Hol+12}: A program that seeks to tackle a major challenge in the life sciences: How to integrate the wealth of data we have acquired post-genome, not just in a mathematical model, but more importantly, in a series of models that are linked across scales to represent organ function. The network is composed of 44 projects, 69 Principal Investigators, and some 200 contributing scientists, distributed widely across Germany. The programme integrates disciplines ranging from molecular and cellular biology, through mathematics, physics and computing, tissue engineering and imaging to clinical medicine.

- The Atlas experiment \cite{Tue08}: A collaboration of 3000 scientists and 174 universities from 38 different countries working on an all-purpose detector for the Large Hadron Collider (LHC)\footnote{http://atlas.ch/}. The detector, which first started collecting data in 2008, is investigating a wide range of physics. Some of the disciplines of its members include high-energy physics, mechanical and...
electrical engineering, semiconductor technology, cryogenics, and computer science.

These examples show a diverseness of collaborations that is not captured in Sonnenwald’s definition; hence scientific collaborations need to be further classified. Sonnenwald [Son07] suggest that collaborations can be divided by their focus; that is, Geographical, Disciplinary, or Organisational. In another example, Shrum [SGC07] classifies collaborations by size.

In the spectrum of scientific collaborations, the size dimension is distinctive because although large collaborations have a significant number of members, there are more instances of small collaborations. Newman [New01] performed a series of bibliographic studies to show that the majority of scientific collaborations have a median number of collaborators in the range of dozens. Guimera et al. [Gui+05] and Lariviere et al. [Lar+14] produced similar studies supporting Newman’s results. Moreover, Liu’s study of collaboration networks [Liu+11] has shown that fewer than 20% of scientific collaborations last for more than one year. Therefore, it could be said that the collaboration spectrum is populated more frequently with small collaborations.

Most of these collaborations are publicly funded, and they are required to share the data publicly (i.e., Data Publishing). Both large and small collaborations have relied on specific infrastructure to perform data-sharing. In the next section, we discuss the information systems used by collaborations to share their data.

2.3 Science Data Repositories

Let’s put the Scientific-Collaboration discussion on hold for a moment and focus on the information systems that scientists use to share their data. In this section, we describe one of the information systems which scientists use to share data and which is also the information system that is the focus of this research: namely, the Science Data Repository.

The Science Data Repository is a web-based information infrastructure that seeks to ensure the storage and accessibility of the research data. Such infrastructures can include so-called data archives, data centres digital libraries, digital collections and the like. Unlike other mediums of data storage and accessibility,
Table 2.2: Different Mediums to Share Data and their features.

<table>
<thead>
<tr>
<th></th>
<th>e-mail</th>
<th>Personal Website</th>
<th>Institutional Repository</th>
<th>SDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistent</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Unrestricted Access</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Web based</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Collaborative Sharing</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
</tbody>
</table>

SDRs support persistent and unrestricted access to data through the web for collaborations; as shown in Table 2.2.

It is complicated to estimate the full spectrum of the information systems used by scientists to share data. There are significant attempts to classify the information systems scientists use to share their data (e.g. Bos et al. [Bos+07], Marcial and Hemminger [MH10]). However, Pamper et al. and the re3data.org project [Pam+13] present the most comprehensive dataset for estimating the diversity of information systems to share data. The project Registry of Research Data Repositories has indexed scientific data repositories from 2013 [Pam+13] and, to date, (March–2016) there are more than 1500 scientific data repositories indexed.

Based on the latter and regarding infrastructure, we know that information systems fall into three categories: (1) repositories based on a data repository development framework, (2) repositories based on a content management system, or (3) ad-hoc repositories. Repository development frameworks are reusable software environments that provide particular functionality as part of a larger software platform in order to facilitate the development of scientific data repositories (e.g. CKAN, DSpace [Lew14] and SEEK4Science [Wol+15]) with specific features for scientific data (e.g. Pre-reserving DOI and Schema cataloguing [Amo+15]). Content management systems are general-purpose development frameworks that designers can customise to create an SDR (e.g. Drupal, Fedora Commons and Eprints [Lew14]). Ad-hoc infrastructures are those defined by the research collaborators themselves; sometimes without taking data management ideas from other systems.

In terms of social-technical organisation, these information systems vary in
two dimensions: type of membership and organisational focus. The membership type is related to the access to deposit assets in the repository. Membership falls into four broad categories: (1) institutional membership, (2) fee-based membership, (3) open archives or (4) registration membership \cite{Pam+13}. One important characteristic of most of the registration-based SDRs is that they use self-deposition as the primary method of data archiving. Self-deposition differs from other approaches in that the data producer, the curator or the data owner performs the data archiving and is responsible for the final decision-making on data accessibility. On the other hand, the second dimension, organisational focus, refers to the domain of the repository, which may be the institution, or the discipline \cite{Pam+13}. In Table 2.3 we presented four examples of SDRs in terms of the dimensions described here.

<table>
<thead>
<tr>
<th>SDR</th>
<th>System Type</th>
<th>Membership</th>
<th>Deposition</th>
<th>Organisational</th>
</tr>
</thead>
<tbody>
<tr>
<td>SysMO SEEK</td>
<td>SDRF (SEEK4Science)</td>
<td>Institutional (SysMO Consortium)</td>
<td>Self-Deposition</td>
<td>Disciplinary (Systems Biology)</td>
</tr>
<tr>
<td>SEEK</td>
<td>SDRF (CKAN)</td>
<td>Institutional (University of Stirling)</td>
<td>Self-Deposition</td>
<td>Institutional (University of Stirling)</td>
</tr>
<tr>
<td>DataSTORRE</td>
<td>SDRF (Dataverse)</td>
<td>Registration</td>
<td>Self-Deposition</td>
<td>Open</td>
</tr>
<tr>
<td>Harvard</td>
<td>SDRF (Dataverse)</td>
<td>Registration</td>
<td>Self-Deposition</td>
<td>Disciplinary (Cellular Biology)</td>
</tr>
<tr>
<td>Dataverse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolights</td>
<td>Ad-hoc</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Examples of Scientific Data Repositories

1. https://www.fairdomhub.org/
2. https://datastorre.stir.ac.uk/
3. https://dataverse.harvard.edu/
4. http://www.ebi.ac.uk/metabolights/

Although the current full spectrum of SDRs has yet to be captured, we can notice some trends arising from the work of the re3data.org project. Concerning infrastructure, 10% of the scientific data repositories were developed using a repository development framework, while the rest were unable to identify the information system architecture that they use. One scenario that could explain this is that the designers used an ad-hoc data repository that did not fit the classification provided by re3data.org \cite{Tea12}. In terms of socio-technical organisation, the majority of the repositories have been self-identified as purely disciplinary (60%) and not affiliated with an institution. These results have been
2.4 Challenges for Small Collaborations

Now that we have defined the core concepts of this thesis, we can return to our motivating example and think of it with regards to these concepts. In essence, we can say that Anna is working in a small geographically distributed collaboration and that she and her collaborators are failing to perform (1) data self-deposition and (2) data publishing using a discipline focused SDR.

Failing at self-depositing data is not unique to our motivating example. Scientists’ reluctance to adopt scientific data self-deposition has been widely reported and highlighted as a critical issue which needs to be solved in the scientific community. Cragin et al. case study [Cra+10] points out that data-sharing outside the inner circle of a scientific collaboration takes place only upon request of the data. In the same vein, Tenopir et al. [Ten+11a] found out that as few as 20% of scientists deposit data in an SDR with the purpose of sharing. These results are supplemented by a follow-up study [Ten+15] which shows that scientists still depend primarily on personal storage options for data-sharing, rather than SDRs. Similarly, Harper et al. study of 12 Biology collaborations highlighted SDR adoption as the main problem for data-sharing success [Har12]. By not adopting self-depositing in SDRs, scientists fail to perform real data publishing and instead adopt other less productive sharing strategies (e.g. Web-posting or sharing upon demand).

Data publishing by users of self-depositing SDRs has not been totally accepted, partly due to the use of access control settings. Designers of scientific data repositories have implemented access and use controls to encourage scientists to self-deposit more data. A survey of data repositories by Eschenfelder [Esc11] (n=175), identified that 20% of the data repositories surveyed used mechanisms to restrict their data collections from the public. For instance, Dryad, a successful repository, includes a temporal data access restriction upon deposition [Roc+14]. Recently, a comparison of SDR frameworks and services, by Amorim
et al. [Amo+15], has shown that access-control is one of the default characteristics of these information systems. This trend could have been fuelled by the access control concerns expressed by scientists. Surveys such as that conducted by PARSE-Insight [KH09] have shown that fewer than 25% of scientists make their data openly available to everyone. Similarly, Tenopir et al. survey study on data-sharing practices [Ten+11a] found that 41% of scientists are not willing to place all of their data into a central data repository with no restrictions. The work of Tenopir et al. [Ten+11a] is complemented by a follow-up study four years later [Ten+15] and Jessel et al. [JB14] study of Figshare, which shows that early adopters restricted the access to 22% of data on Figshare. This view is supported by case studies such as that conducted by Cragin et al. [Cra+10] and the Research Information Network (RIN) [RIN08] which found that scientists often need to restrict, some or all, of their data from public access for some length of time. The increasing number of SDRs that allow data scientists to set restrictions on the data they self-deposit, coupled with a prevailing “controlling” culture in science has potentially given rise to a trend of restricting access to datasets in SDRs.

These two challenges, Scientists’ adoption of self-deposition and scientists’ adoption of data publishing, stand at the core of the problem of scientific data-sharing at large, and they have been the focus of a significant number of efforts.

2.5 Coping with Data Withholding

To address this issue, the community has brought up some solutions, which are categorised in two themes, these being Mandates and Incentives.

Policy and Mandates

One of the primary mechanisms used to address data withholding has been the implementation of data-sharing policies. Data-sharing policies are conditions attached to funding schemes in order to encourage researchers to share their data [RC08]. Ruusalepp’s study of data-sharing initiatives has shown that these kinds of policies are widespread in OECD countries [RC08]. However, these policies are not exactly mandates but intention statements [RC08] [MP13]. Data-sharing mandates, on the other hand, have been established by institutions such as journals. Numerous journals now require authors to share their raw data sets as
a condition of publication [SV09]. For example, in their survey of data-sharing policies (n=371) Sturges et al. [Stu+14] found that as many as 50% of the journals had some sort of data-sharing policy in place, and nearly 40% of them indicate that data must be deposited in some information system.

**Incentives**

In order to incentivise data-sharing, the community has resorted to creating new incentive mechanisms. Data are being considered as scholarly resources and thus can be cited. A Force11 project has formed a working group to develop guidelines for implementing the joint data citation principles for direct data citations in scholarly articles [Cla14]. Besides, different types of recognition mechanism are being implemented for data. The Data Usage Index [IC11] has been developed to recognise the impact of biodiversity datasets by measuring their use by third-party user communities.

The solutions to the data-sharing problem presented by the community have focused on providing incentives and mandates (in other words, “carrots” and “sticks”). Nevertheless, from a risk-cost-benefit perspective, it could be argued that the current perceived risks and cost derived from data-sharing outweigh the scientists’ perceived benefits.

A number of published studies have described the variety of costs that scientists may pay when sharing their data. In a study which set out to identify data-sharing practices, Tenopir et al. [Ten+11a] found that the time spent in preparing data for sharing and the lack of funding to account for this preparation were among the major costs which scientists have to pay when sharing data. Jessel and Birch [JB14] obtained similar results in their survey of early-adopters of Figshare and, more recently, a follow-up survey of the Tenopir et al. study shows that these costs are still perceived [Ten+15]. These costs are important considerations, as in an SDR using self-deposition, scientists have to bear all these costs individually. These costs are significant barriers to adoption.

Similarly, there is evidence that highlights the perception of inherent risk in sharing data in an SDR. Scientists, depositing and making their data accessible in an SDR, give away control over their datasets to anyone with access. Most of the available surveys (e.g. [Ten+11a], [Esc11], [JB14] and [Ten+15]) on data-sharing attitudes have identified scientists’ perceived risk in the form of fear of misinterpretation – i.e., using the data in a way that was not intended. The
misinterpretation of results could have unintended consequences that could damage the dataset producer’s reputation. Case studies in specific fields of Physical Sciences have also found out that scientists expressed fear (i.e., perceived risk) of being scooped by competitors \cite{Cra+10, Vel13}. Scooping could hinder the ability of scientists to pursue further funding or work opportunities. These risks (i.e., potential misinterpretation and scooping) together with costs are aspects that have developed into adoption barriers for SDRs.

Data-sharing can benefit both the scientific community as a whole, and individual scientists. However, the latter have not expressed that they perceive individual benefits. The community has foreseen three potential individual benefits: (1) Data citation from third-party dataset reuse, (2) open data citation boost, and (3) co-authorship from third-party dataset re-user. Data citation from third-party dataset reuse, probably the primary individual benefit for a scientist, has not yet been perceived to be a real benefit by the community \cite{Nel09, Fer+14}. Mooney et al. study on data citations practices \cite{MN12} has shown that the majority of articles failed to include an adequate citation to secondary data. Similarly, Alsheikh-Ali et al. study on published scientific data availability \cite{AA+11} found that, in the majority of cases, data citation practices in publications are not done in a manner consistent with the policies of the publishing journal. Although alternative attribution mechanisms have been proposed (e.g. ALT metrics \cite{DAD14} and Data-Usage-Index \cite{IC11}) none of them have been fully adopted by the scientific community. This lack of adoption and support for data citation mechanisms is one of the barriers that hinder the possibility of the perception of data citation as a tangible benefit. The latter is more apparent when one considers that data citation has been identified as the main requirement for data-sharing engagement in a number of surveys (i.e., \cite{Ten+11a, JB14, Ten+15}). Another individual benefit is what has been termed as “open data citation boost”. Open data citation boost refers to the phenomenon in which articles associated with deposited materials receive a boost in citations after data deposition has occurred. Boost in publication citations has been linked to the open publication of datasets (e.g. clinical trials \cite{PDF07, PV13} and biomaterials \cite{FS06}). This mechanism alone should be a good incentive for scientists to share their data, as publication citation credit is well supported in academia. However, tracking the citation boost relies on good data citation practices that currently are not in
place, as shown by previous studies (i.e., \cite{AA+11}, \cite{MN12} and \cite{PV13}). Additionally, with the exception of a case study conducted by RIN in 2008 \cite{RIN08}, we found no evidence in surveys that citation boost is perceived as a potential benefit (e.g. \cite{Ten+11a}, \cite{JB14} and \cite{Ten+15}). Finally, co-authorship from third-party dataset reuse refers to the emergence of collaboration when a third-party identifies a shared dataset worthy of further analysis. Despite the fact that this has not been pointed out as a benefit or a motivation to share data, it has been expressed as a secondary expectation from third-party dataset reuse. For example, co-authorship was emphasised by many scientists in the case study from Cragin et al. \cite{Cra+10}. It ranked second as a recognition requirement from scientists in the surveys conducted by Enke et al. \cite{Enk+12} and Tenopir et al. \cite{Ten+11a}. One of the potential reasons behind the lack of perception of this as a benefit is that authorship agreements, in these circumstances, do not appear to be a regular part of the scholarly process \cite{Cra+10}.

The perception of risk, presence of cost and lack of knowledge of individual benefits is the context in which data-sharing is developing.

We argue that, regardless of the fact that the community has been trying to address this context through mandates, incentives and infrastructure, the solutions have not trickled down to the methodologies for designing Scientific Data Repositories. Specifically, we argue that to fight these challenges, the SDR’s designers need to acknowledge the context of data-sharing while simultaneously considering the restrictions imposed by the characteristic of the collaborations around the scientists – which in most cases are small collaborations. This argument has been echoed by a number of research efforts in terms of:

- Considering the roles of data \cite{BB03}.
- Acknowledging data practices in their user communities \cite{Bor07}.
- Considering the context of data production \cite{VD11}.
- Taking into account the motivational aspects at all stages of data submission \cite{Enk+12}.
- Considering field-inherent sources for differences in openness and sharing behaviours \cite{Vel13}.
- Considering the context \cite{Hae+13}.
Table 2.4: Comparison of approaches to capture and incorporate context into SDRs. Here we compare the different documented approaches in the literature to design SDR. Additionally we provide a set of examples of SDR builds using that approach.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Context Capturing</th>
<th>Context Incorporation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Lin et al., 2008)</td>
<td>Experience*</td>
<td>Scenario-Based</td>
<td>myExperiment</td>
</tr>
<tr>
<td>(Wolstencroft et al., 2011)</td>
<td>Experience*</td>
<td>Adaptive</td>
<td>SysMO SEEK</td>
</tr>
<tr>
<td>(Rosson et al., 2012)</td>
<td>Socio-Technical</td>
<td>Scenario-Based</td>
<td>HRCT</td>
</tr>
<tr>
<td>(Lassi, 2014)</td>
<td>Socio-Technical</td>
<td>Scenario-Based</td>
<td>NA</td>
</tr>
<tr>
<td>(Myers et al., 2014)</td>
<td>Experience</td>
<td>Scenario-Based</td>
<td>BioenergyKDF</td>
</tr>
<tr>
<td>(Lagoze et al., 2014)</td>
<td>Experience</td>
<td>Scenario-Based</td>
<td>CE2DAR</td>
</tr>
</tbody>
</table>

* Software Design for Empowering Scientists (De Roure and Goble, 2009)
NA: Not applicable

- Acknowledging discipline- and individual-level factors [KS15] [Kim13].
- Acknowledging data cultures [TP11] [Ten+15].

Therefore, we do not propose a new type of Scientific Data Repository but a design approach to design SDRs which more effectively incorporate the context of data-sharing. In the next section, we compare and contrast previous research and practices on SDR design that highlight the existing gaps in incorporating context into the SDR infrastructure.

### 2.6 Related Work

In this section, we review the related work in the area of SDR design. In the work ahead, we have identified the components required for designing an SDR: Context-capturing and context incorporation. We use these components as a framework to compare existing SDR design solutions for small collaborations, and we ground them in the SDR development Life Cycle. The Table 2.4 summarises the SDR design solutions outlined in this section. These solutions address each of the components we have outlined above in different ways. Where the authors have not described in detail the mechanism used for part of their approach, we assume that it was achieved through bespoke practices.
2.6. RELATED WORK

2.6.1 SDR Development Life Cycle

To address the challenges previously presented for development, we ought to examine the design and development of different Scientific Data Repositories. Before discussing the way that others have addressed these challenges, we define the SDR design and development terminology that will help us to discuss the comparison in the next section.

![SDR development lifecycle. Highlight the stages for context incorporation](image)

Figure 2.1: SDR development lifecycle. Highlight the stages for context incorporation

The SDR development life cycle follows the standard model of System development life cycle; which is shown in Figure 2.1. This is an iterative process; that comprises six stages: Planning, Research, Design, Adaptation, Measuring and Release. All the stages are important, but we focused on the Research and Design stage as they present the most creative parts of the process.

The Research Stage, is the second stage of the SDR Development Lifecycle. In this stage, the designer collects information to ensure the design employs a variety of methods and information sources. In the case of SDRs, these information sources are (1) users and (2) context. User Requirements Capturing consists of employing several techniques to elicit the requirements from the users and creating requirements lists. Context-capturing consists of the acquisition of contextual information about the users’ activities and their motivations that aid the development of design. There are two standard approaches for context-capturing: (1) experience-based approach and (2) the formalised socio-technical approach.

The Design Stage is the third stage of the SDR Development Lifecycle. In this step, the features meant to prompt a particular behaviour are defined. The definition of these features follows two practices: Adaptive or Scenario-based. The Adaptive practice consists of constructing a design feature based on a design feature from a different system that prompted the desired behaviour. For instance, when designers use an SDR development Framework (e.g. DSpace,
CKAN, DataVerse) one could say that they are adopting features from other designs. Under the **scenario-based** practice, designers define design claims based on task analysis and a scenario of use.

### 2.6.2 Context Capturing

Most current SDR implementations rely on gathering contextual information of the scientific endeavour through two approaches: (1) the designers experience or (2) using a formalised socio-technical approach. For example, Wolstencroft et al. [Wol+11b] and Li et al. [Lin+08] based context inclusion into the design of the SEEK (a catalogue for Systems Biology materials) and myExperiment (a repository of scientific workflows) in six principles that they synthesised over years of developing information systems for scientists [DRG09]. Similarly, Lagoze et al. [Lag+14] and Myers et al. [Mye+14] based their inclusion of context in their design of the CED2AR repository (a metadata repository for the Humanities) and the BioenergyKDF repository (a Biodiversity repository) in the years of experience they have gathered. On the other hand, a limited number of approaches use a formalised approach in order to bring context into the SDR design. These approaches consider that SDRs are socio-technical platforms. For instance, Lassi and Sonnenwald [Las10] used a taxonomy that provides the designer with an overview of explanatory factors concerning adoption and use of systems such as SDRs that can be used to design casual claims about design features. They used an empirical study with their potential users to inductively create the taxonomy. Another example is the work of Rosson et al. [Ros+12]. In this case, they brought in contextual information to create claims from a case study of the community they were designing for – a project that studies primate morphology and palaeontology with gene mapping in baboons and mice.

Although both approaches bring context to the design, there is still scope for improvement. The experience-based approach’s biggest problem is related to the designers’ experience. This experience could be (1) unrelated to the data-sharing context and (2) not up-to-date with empirical findings of affecting data-sharing. This first problem is often encountered in other types of application in which the context is important. For instance, LeRouge et al. [LeR+13] argues this is the case with designers of health applications for the elderly. In these applications, design and development are typically done by younger, presumably, healthy adults who have a tendency to put in their biases by designing for themselves and
not necessarily for the elderly users. Regarding knowledge keeping up with the latest findings, the experience-based approach needs to be updated with the growing body of literature on data-sharing. For instance, In recent years, there has been an increasing amount of literature on scientific data-sharing (e.g. [Piw11], [WBM11], [KFH12], [KS15], [FFH15]) and it would be naïve to expect designers to keep this in mind if a formalised methodology were not in place. Thus, there is a methodological conflict in that the experience approach to context-capturing is unable to guarantee related context-capturing and sufficient context to capture.

In the case of the formalised socio-technical approach as presented in the literature, by Lassi [Las10] and Rosson [Ros+12], we identified one concern. The designer needs to find a collaboration, working with an SDR, in order to be able to capture context. In other words, the context is captured a posteriori of the data-sharing experience. This has been a criticism expressed to this empirical approach by Carroll [CR92]. Recently, Kraut [Kra] has drawn a parallel criticism, arguing for an analytic approach to the design of online communities. In essence, the formalised socio-technical approach would be improved if additional sources of context information could be integrated into this stage.

To complicate matters even more, the fact that most of these design approaches focus on small scientific collaborations, limits the range of research methodologies used with either of the approaches. Capturing behaviour related to context becomes a complicated issue as SDRs could be classified as an online peer-production system and they follow a power-law distribution regarding depositions. Online peer-production systems are self-contribution or peer production systems in which groups of individuals collaborate to provide digital goods. The product of this collaboration can be considered a commons [BN06]. The contributions in these systems are overall characterised by a power-law distribution which means that a handful of distinctly active users account for most of the depositions, and most of the users make no depositions [Vos05] [Stv+08] [Wil08]. Therefore, context-capturing for small collaborations, which have a finite pool of contributors, is limited to certain methodologies that limit the possibility of identifying statistical relationships with self-depositions (i.e., qualitative methodologies).
2.6.3 Adapting Strategy

One common practice among SDR designers to incorporate context is the adaptation of admired design features from similar information systems. The results of this practice in SDRs have been mixed and thus the practice is not always reliable. In the SEEK platform, for instance, “discussion forums” were included under the design claim that “they would allow users to share knowledge with each other” [Owe10]. That feature was later deprecated due to a lack of use [Owe15]. On the other hand, one feature that has been adapted into most modern SDRs is Digital Object Identifier (DOI) assignation. A DOI is a permanent, globally unique identifier that is a major component of data citation. One of the first SDRs to include DOI assignation was Dryad [Car08] and, overtime, this has been adapted to many other SDRs to the point that Scientific Repository Frameworks (such as DSpace) and Scientific Repository cloud services (such as Figshare and Zenodo) now provide a pre-reserving DOI feature [Amo+15].

Adapting design features from other platforms can be seen as a shortcut to reduce time and effort at the design stage. However, as exemplified above, there is no guarantee that adapted design features cause the desired behaviour. Previous work has draw parallels between this problem and the phenomenon of “motivational misalignment”. Motivational misalignment occurs when the subjects’ motivational profile is affected by an intervention with opposite effects of that desired by the designer [Ben]. For instance, Jan et al. [AJ11] has shown that employees who are assigned to edit entries in corporate-adapted wikis have lower participation than those employees who make wiki-edits for enjoyment. This change in participation could be caused by a misalignment between the types of incentives a user has to perform an activity. If both incentives are of the same type, a user would be more likely to earnestly participate in that activity (e.g., “I enjoy making wiki-edits” and “my participation is enjoyed by others”). On the other hand, if the incentives are of different type (e.g., “I enjoy making wiki-edits” and “It’s my job to do wiki-edits”) that might have opposite consequences than those intended. In their work on designing online communities, Kraut and Resnick [RK12] argue that motivational misalignment is one of the biggest problems with adapting successful practices (e.g. the use of wikis). Therefore this poses a methodological conflict in that a modification of the methodology needs to be proposed in order to deal with the reliability issues that could be potentially created due to motivational misalignment.
2.6. RELATED WORK

2.6.4 Scenario-Based Strategy

The other prevalent practice for creating design features is the scenario-task strategy. In this strategy, designers build scenarios of tasks around which design features are built \[\text{CR92}\]. The empirical approach to scenario building consists of observing people and asking them what they do \[\text{CR92}\]. This scenario building work allows designers to incorporate context into the design features. For example, the High-Resolution Computed Tomography Repository (HRCT) \[\text{Ros+12}\] provides a feature that enables users to add rich metadata (e.g. investigator, project, specimen, the source of a specimen) about the dataset deposited in the HRCT. The task under which the feature was designed comes from a real scenario (i.e., Note taking in a paper-based lab-notebook). The design claim is that having the ability to add metadata in electronic form would cause the users to migrate from the paper-based lab-notebook to the electronic form.

Although this strategy can bring context to the design, the degree of context incorporation is dependent on a rich context-capturing stage, which is not always guaranteed in the case of the context of data-sharing. As shown in the previous section, some surveys have exposed that data-sharing is not a task commonly performed by the scientific community (e.g. \[\text{Ten+11b}, \text{KH09}\]). Therefore this poses a methodological conflict, in that a modification to the methodology needs to be proposed in order to deal with the limitation of purely relying on a sound context-capturing stage whilst still incorporating sufficient context into the design features.

2.6.5 Gaps

In summary, we identified four methodological conflicts\[^\text{6}\] that need to be addressed in the state of the art of Scientific Data Repositories design methodologies.

**Gap1:** First, the experience-based approach is unable to guarantee related context-capturing and sufficient context to capture. See sub-section 2.6.2.

**Gap2:** The socio-technical empirical approach to context-capturing

can only capture the context in a-posteriori fashion. See subsection 2.6.2.

**Gap3:** The adapting strategy to context incorporation has reliability issues which could potentially be caused by motivational misalignment. See subsection 2.6.3.

**Gap4:** The degree at which the scenario-based strategy incorporates context to design is limited to sound context-capturing practices. See subsection 2.6.4.

In the next section, we propose two modifications to current approaches to SDR design to address these methodological conflicts.

### 2.7 Contribution

In this work, we proposed an alternative approach to incorporate the context of data-sharing and address the gaps identified in the SDR design approaches and practices.

#### 2.7.1 Research Stage

As outlined in previous sections, both of these context-capturing approaches have downsides, but the formalised socio-technical approach provides an upgrade over the experience-based. The socio-technical approach allows designers to construct a data-sharing model of the collaboration and, using this model, they can identify the data-sharing Context. In this sense, the approach changes from being deductive to inductive and potentially leads to a better fit with the collaboration’s data-sharing Context. For instance, by using the socio-technical approach, designers would need to provide empirical evidence rooted in the case at hand; thus they would be less inclined to propose design features that rely on previous experience. Thus, concerning context-capturing, we propose to address the issues related to capturing the related and sufficient context in the experienced-based approach (**Gap1**) by using a socio-technical alternative such as that proposed by Lassi [Las10] and Rosson [Ros+12].

For addressing the problems with the a-posteriori context development (**Gap2**), we propose complementing the socio-technical approach with the use of a theory-inspired approach that allows practitioners to bring data-sharing context from
other cases with similar data-sharing models. Specifically, we propose defining this approach on the basis of the Translation of Social Theories to interventions approach as outlined by Kraut and Resnick [RK12]. Kraut and Resnick use this approach to design online communities. For example, Ren et al. [Ren+12a] used theories of identity-based attachment from Social Psychology, to increase member attachment to a movie-related online community. They gave community members information about group activities and intergroup competition, and tools for group-level communication. In other contexts (i.e., not online) members’ attachment has been shown to increase when information about the group and communication is facilitated. Kraut and Resnick argue that, through the theory-inspired approach, they can provide a practical lens through which designers can look at their decisions in a systematic fashion, rather than using overly general experience or through trial and error. Moreover, they argue that the theoretical insights supported by empirical evidence are powerful tools that designers could leverage to build vibrant online communities. This approach could allow designers to begin the research stage with a starter template, rather than to rely solely on a-posteriori context-capturing from their case. Therefore, using an empirical context-capturing approach complemented with an analytic approach for capturing context (as shown in Figure 2.2), designers can avoid bias imposed by previous experiences or lack thereof (Gap1), and they can capture context that would not be available a priori (Gap2).

Figure 2.2: This thesis’s socio-technical method using an analytic approach to context-capturing. This approach complements the empirical strategy used in the literature.
2.7.2 Design Stage

In terms of incorporating context to the design stage, we propose to address the reliability problems of the adapting strategy (Gap3) and the limitations to incorporating context of the scenario-task strategy (Gap4) by using Persuasive Design as the main design approach for defining design features. Specifically, we propose a procedure to generate design claims using Choice-Architecture [TS08]. Choice-Architecture is a Persuasive Design Strategy based on the idea that changes in the decision environment can affect individual decision-making and behaviour while preserving freedom of choice. The strategy alters people’s behaviour in a predictable way without restricting options or significantly changing their economic incentives. To illustrate Choice-Architecture in a more general setting, consider the example of retirement plans in the US. Employers offering retirement plans are responsible for selecting the plan’s menu of investment offerings, but employees retain ultimate responsibility for their portfolio construction. However, employees are often ill-equipped to make such decisions. The retirement savings marketplace has answered this concern with the introduction of target-date funds retirement plans. These funds seek to simplify employee investment decision-making. Target-date funds offer three distinct portfolio features to investors. For instance, one such feature is an age-based rebalancing service, whereby risk exposure is reduced automatically over time. The rebalancing can be viewed as a form of commitment device that the employee might otherwise fail to implement on his own. It has been shown that this “Choice-Architecture” is accelerating demand for target-date funds [MU12]. In this way, the implementation of Choice-Architecture managed to alter employees’ behaviour in a predictable way without forbidding any options.

The example of retirement plans might show a clear implementation of Choice-Architecture, but not all decision-making problems are suited to use the Choice-Architecture method. Its reliance on cognitive biases or heuristics only makes Choice-Architecture applicable to problems where the decision is (1) complicated, and (2) infrequently made. Additionally, Choice-Architecture is applicable to decisions in which (3) the benefits of the decision are delayed, but the decision is born with cost, (4) the target-group has poor feedback about the decision’s outcomes, and (5) the decision outcomes are difficult to imagine. When these five requirements are met, one can use Choice-Architecture [TS08].
We argue that the data publishing decision problem in science is a suitable challenge for the Choice-Architecture paradigm. To demonstrate the latter statement, consider the following assertions from the literature. First, Choice-Architecture applies to problems where the decision is complicated. Similarly, the decision of archiving data is made by individual scientists \cite{Ten+11b} and, at an individual level, scientists perceive associated costs, benefits and risks with open data-sharing that add complexity to the decision-making process \cite{Roc+14}. Second, Choice-Architecture is fit for decisions made infrequently. Likewise, note that data-sharing decisions are made after a study is published and that scientists publish with a median below three papers per year \cite{NP15}. Thus, the frequency at which scientists take the decision is effectively low. Third, the paradigm is suitable for decisions that are born with costs and have delayed benefits. Similarly, all costs associated with data-sharing are to be borne around the publication stage \cite{Ten+11a}, while all benefits are to be borne some time after publication; creating a considerable lapse between cost expenditure and the observation of benefits. Finally, Choice-Architecture is suited to decisions in which the target-group have poor feedback about the outcomes, and in which the outcomes are difficult to see. Likewise, the potential sources of feedback for open data-sharing – (1) data citation and (2) “data citation boost” – are currently not fully supported \cite{MN15} or lack of tracking mechanisms \cite{PV13}. Mooney and Newton \cite{MN15} have shown that secondary data citation is not a common practice in different fields of science. Similarly, Piwowar and Vision have argued that poor citation practices hinder the effective tracking of data citation boost \cite{PV13}. Thus, we argue that the data-sharing decision problem fits the Choice-Architecture paradigm. Therefore, by designing features under the Choice-Architecture approach, we depart from the reliability problems of the adapting strategy (\textbf{Gap3}) and we directly model context into the design features without completely relying on a rich context-capturing stage (\textbf{Gap4}). Figure 2.3 shows a high-level representation of this strategy.

Overall, we argue that, in order to create effective Scientific Data Repositories which address self-deposition and data-sharing, context needs to be added at each stage of the SDR development life-cycle. Moreover, we argue that there are methodological conflicts in the way that current approaches carry out the data-sharing context incorporation to the design features of the Scientific Data
Repositories. We propose addressing these methodological conflicts with an analytic approach to context-capturing and the use of a procedure in which to incorporate context in design features based on Choice-Architecture.

**Related Work on Persuasive Design**

Choice architecture is not the only persuasive design approach that we could have used but the model underpinning the approach offers a good fit for the data sharing dilemma. First, at one level, Choice Architecture is based on a contextual model of behaviour change and thus is a fitting model for instances when interventions are focused on the context surrounding individuals. At a second level, unlike other contextual models, the Choice Architecture model is useful to make predictions about decision-making problems such as the data sharing dilemma because of its grounding in psychology and its ability to model infrequent behaviour.

Nierdderer et al. [Nie+14] classifies behavioural change models and approaches in two big groups: Individualistic and Contextual. The individualistic models focus on individuals’ agency. They are founded on three principles: (1) choices are rational; (2) the individual is the appropriate choice of analysis; (3) behaviours are self-interested. Examples of these models include the Health Behaviour Model, Theory Planned Behaviour [Ajz91], Stages of Change, and Behavioural Economics. Examples of design approaches under these models include the Design for Healthy Behaviour approach, The Loughborough model, the Behaviour Grid, and Persuasive Technology [Fog02]. On the other hand, the contextual models suggest that behaviours are consequences of social norms and expectations held in place by the context in which individuals live. Examples of these models include
the Christmas Change Model, Social Practice Theory and Choice Architecture. Examples of design approaches under these models include Product-Impact tool, Moralised products and MINDSPACE (i.e., a subset of Choice Architecture). One of the main differences between the two model categories is the focus of the unit of analysis; the individualistic models use the individual, whilst the latter use the context. At the same time there is significant overlap on some of the models; specifically Behavioural Economics and Choice Architecture.

In this work, we choose the contextual models as we aim to explore the effects of technologies used by individuals in their data sharing decision making. These technologies are effectively their context. Additionally, we discarded approaches in which the underpinning model was not originated in psychology [Nie+14]; that is the case for the Social Practice Theory model [Dor12].

In terms of which contextual design approach is better fit for the data sharing problem, we identified a few differences between the Moralised products approach [Jel06] and MINDSPACE [Dol+12] (i.e., a subset of Choice Architecture [TS08]). Both approaches have been successfully used in a multitude of scenarios. However, the moralised products approach has consistently been used for routine practices and, as we have shown at the beginning of this subsection, that is not the case with the data sharing practices. Data sharing in all explored scenarios is a non-frequent activity.

2.8 Summary

In this chapter, we have presented the state of the art approaches to design Scientific Data Repositories. We grounded the work of this thesis in the SDR Development Life Cycle and within the demographic of small collaborations. In place, we identified four methodological conflicts or gaps in the current approaches to design SDRs and proposed two methodological changes to address these gaps.

In the rest of this thesis, we detail and apply this proposed approach to incorporate context into design features in an SDR used by a Small Collaboration from the Life Sciences, and we evaluate the efficacy of the design features. We structured the application of the approach and its evaluation by chapter, as shown in Figure 2.4. Chapter 3 presents the application of the context capturing stage to a small scientific collaboration from the Life Sciences and the study that evaluates of the captured context. Chapter 4 presents the application of the
context-incorporation stage to an SDR used by a small collaboration from the Life Sciences, and the study that evaluates this part of the approach.

Figure 2.4: Organisation of the application of the Persuasive Design Approach in terms of the chapters ahead.
Chapter 3

Translating Theories

They know you can do anything.
So the question is: what don’t you do?; rather than what do you do?

David Fincher

Previously, in Chapter 2, we identified two variants of the socio-technical approach as candidate approaches to address context capturing in the Scientific Data Repository development life cycle. In this chapter, we present an investigation of the application of this thesis’ context-capturing approach to a set of small scientific collaborations using an SDR.

The application of this thesis’s approach to context-capturing was conducted in three stages: exploration, review, and translation. We apply this context-capturing approach to a set of small collaborations from the field of Life Sciences. First, we conducted the exploratory study to capture context inductively. The results of this study allowed us to (1) identify the constructs we would use in the translation study and (2) to identify a set of factors associated with data-sharing in an inductive fashion. Then, we carried out a review of factors associated with data-sharing in the specific field of Life Sciences. Finally, we assessed the feasibility of translating the factors from the review. Figure 3.1 shows the studies we conducted in the context of this thesis’s proposed approach to capturing the context of data-sharing.

We found that the application of the analytic context-capturing approach in a small scientific collaboration proved to be more complicated that expected. First, empirical factors to be tested deductively are of an archetype that cannot
be used as an intervention; that is, these factors are characteristics of the cases (e.g. composition of the collaborations). Secondly, the size and life span of the collaborations made it more complicated to draw concise conclusions, even when different types of data were gathered. Finally, we did not manage to engage the full range of collaborations; in other words, critical cases were not included in the study.

Nevertheless, the inductive context-capturing approach allowed to identify a number of factors associated with data-sharing that, in conjunction with the factors identified in the review, we can use to define design claims for an SDR. We found that small collaborations from Systems Biology align their data sharing practices with their collaborators publishing and data lifecycle. Additionally, we found that the collaborators’ data sharing perceptions differ from their actual behaviour. Concerning the data sharing attitudes, we found that collaborators are only extrinsically motivated to share data. They do not perceive any individual benefits from data sharing and the risks they perceive are similar to those in larger collaborations and other fields of science. Altogether, these findings contribute to the body of knowledge of data-sharing in small collaborations and serve as the context for the application of the approach proposed in this thesis to design Scientific Data Repositories.

This chapter begins by presenting the exploratory case study we conducted in order to understand a small scientific collaboration and identify factors associated with data-sharing using the empirical approach to context-capturing. Additionally, we discuss the rationale behind the selection of the single case study as the research approach for this study. After that, we present a literature survey of factors associated with data sharing. We then, discuss the application of the analytic context-capturing approach to test factors in the case of small collaborations using an SDR.

### 3.1 Exploratory Case Study

In this section, we present the exploratory study which we used to understand a small scientific collaboration in the life sciences. This study had two objectives. First, to identify the constructs that could be employed in the application of the analytic context-capturing approach in section 3.3. Second, to identify factors associated with data-sharing that could be used to define design claims in the
Figure 3.1: Empirical and Analytic approaches to context-capturing. We present the three studies presented in this chapter in the context of the approach they serve.

design stage of the SDR development life cycle.

To identify the particularities of the small collaboration, we used an exploratory case study methodology. The case study followed an embedded case study design and we collected data via semi-structured interviews and the SDR’s logs. In the study, we explored three topics: the scientific collaboration process, the collaboration’s data-sharing practices and attitudes.

We found that, on the subject of collaboration, the collaborations suffered the common problems of geographically-distributed collaborations. Regarding data-sharing practices and attitudes, we found that extrinsic incentives motivated collaborators. Collaborators showed a limited awareness of potential benefits and data sharing only occurred under low-risk scenarios. The findings of this study have a number of important implications for SDR designers and scholarly communication researchers of small scientific collaborations.

We begin this section with the description of the exploratory study. We then present the results and a discussion of this exploration. Additionally, we present alternative studies to the case study that we contemplated conducting or that we piloted.

### 3.1.1 Research Questions

In essence, we used this study to analyse a group of small collaborations using an SDR to answer the following research questions:
**RQ1.** What are the processes of collaboration in a small scientific collaboration?

**RQ2.** What are the data-sharing practices and attitudes adopted by scientists in small scientific collaborations?

In the next section, we present the study design we used to answer these two questions.

### 3.1.2 Design and Objectives

We used an exploratory embedded case study design to complete the previously mentioned objectives. The exploratory case study is an empirical strategy used to investigate phenomena characterised by a lack of specially formulated hypotheses that can be tested, or a specific limit on the choice of methodology \[\text{MDW10}\]. An embedded case design is a strategy used to investigate a phenomenon in their context in which multiple units of analysis are studied within the case \[\text{RH09}\].

We selected this design for four reasons; rich data collection, studying an ill-defined phenomenon, identify new questions; good reliability. First, case study design uses a mix of qualitative data collection techniques to be incorporated into the overall research design \[\text{Yin13}\]. These included observation, systems logging data, pre- and post-study questionnaires and interviews. Therefore, adopting these data collection techniques facilitated the compilation of a richer set of data sources than could have been attained by other means. Moreover, in this research, we aimed to focus on data collected from small groups under their natural environment. An examination of the literature suggested that this type of data, when analysed using qualitative methods, could provide valuable new insights \[\text{Yin13}\]. We used an exploratory approach because the phenomenon of scientific collaboration and its relationship with data-sharing is ill-defined. The exploratory case studies are, by definition, often applied in a research context that is not clearly specified and still requires data for the formulation of valid hypotheses; their broad concept provides the researcher with a high degree of flexibility and independence concerning the research design as well as the data collection \[\text{MDW10}\]. Moreover, we used this study as a preliminary step of an overall explanatory research. The exploratory case study design is often applied in a type of research in which the research questions have either not been clearly identified/formulated nor has the data required for a hypothetical formulation been yet obtained.
Finally, we chose embedded case study design in order to provide a more detailed level of inquiry. In the embedded case study design, sub-units of analysis are identified and their differences enhance the reliability of the case study \cite{RH09, Yin13}.

### 3.1.3 Case Selection

The embedded case study design requires the use of multiple units of analysis and thus we used a set of five small collaborations using an SDR that shared the same context. We made use of the following criteria for the selection of embedded cases. First, cases have to be small collaborations as defined in Chapter 2 (e.g. Collaboration on Diagnosis of Emerging Viruses (CoDEV) \footnote{http://www.nottingham.ac.uk/codev/about.aspx}). That means small scientific collaborations with a short life span. Second, all the embedded cases must come from the same discipline cohort. For this study, that is Life Science scientific collaborations. Third, all the embedded cases should use a similar SDR. Fourth, all the embedded cases should share a similar context in the form of similar norms, geographical diversity and institutional diversity.

In order to identify collaborations, we use a convenience sampling method. The convenience sample being: scientific consortia. This method of search allowed us to identify well-defined collaborations even when they were small and had a brief life span. Following the previously mentioned criteria, we selected six collaborations from the Galaxy Groups Programme\footnote{We anonymized the names of the programme and the collaborations in this study using the Galaxy Groups (GrG) names from https://en.wikipedia.org/wiki/List_of_galaxy_groups_and_clusters} as this study’s embedded cases. The Galaxy Groups Programme is a real case from the field of Systems Biology. The programme is an international consortium that uses an online platform to share collaboratively and publicly. The Galaxy Groups Programme is divided into 15 collaborations, each of which assembles an international collaboration to study a different microorganism. Each project has explicit coordinators, who are responsible for coordinating maintenance tasks and keeping the project functioning. In each project, there are liaison members who advocate and communicate the scientific data-sharing policies.

The Galaxy Groups Programme uses the SEEK platform \cite{Wol+11b} to share collaboratively and openly. SEEK is an SDR developed using a Science Data...
Repository framework (SEEK4Science). It uses a type of semi-institutional membership; all members are part of a consortium but come from different institutions. Nobody can subscribe or pay to contribute to the SDR. It uses self-deposition, and it has a disciplinary focus; that is, all the users of SEEK are from the field of Systems Biology.

Table 3.1 shows that the six collaborations we selected (1) use the same SDR; (2) are bounded by the same data-sharing policy; (3) they have similar geographical and institutional diversity; (4) their size and life span matches that of a small collaboration. Additionally, all embedded cases had individual members promoting data-sharing within the collaboration.

Table 3.1: Embedded cases descriptive information. Here we compare and contrast the six collaboration that were studied in this work.

<table>
<thead>
<tr>
<th>Collaboration</th>
<th>DEER</th>
<th>ZWICKY</th>
<th>BULLET</th>
<th>BURBIDGE</th>
<th>COPE</th>
<th>LEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical Diversity</td>
<td>0.22</td>
<td>0.27</td>
<td>0.27</td>
<td>0.26</td>
<td>0.12</td>
<td>0.29</td>
</tr>
<tr>
<td>Institutional Diversity</td>
<td>0.29</td>
<td>0.17</td>
<td>0.21</td>
<td>0.18</td>
<td>0.14</td>
<td>0.22</td>
</tr>
<tr>
<td>Size</td>
<td>7</td>
<td>11</td>
<td>17</td>
<td>12</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>Life span</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

All the collaborations using the SEEK4Science platform as SDR.
All the work under the same Data Sharing Policy are work under the umbrella of the Galaxy Groups Consortium.
For aesthetics we abbreviated the name of two collaborations here DEER:DEER-LICK and COPE:COPELAND.
Lifespan is given in terms of years.
Diversity was calculated in terms of GINI index.
Size is given in terms of number of collaborators.

### 3.1.4 Data Collection

We used direct and independent data collection techniques as the main data collection for this case study. Direct techniques meant that we were in direct contact with the participants and collected data in real time [RH09]. Independent analysis refers to the use of already compiled data, such as documentation [RH09].

**Main Data Collection Instrument**

We used semi-structured interviews as the main instrument of direct data collection. Semi-structured interviews are characterised by their flexibility in the way the questions are asked and replied to. In this type of interview, we had a list of
questions to be covered (i.e. the interview guide), but we noted any additional things which were said by the interviewee. Using semi-structured interviews, we were able to determine the subjects considered to be important. A copy of the interview guide can be found in Appendix A.

We followed a two-step approach for selecting items to be included in the interview guide. First, we reviewed the literature on scientific collaboration and data-sharing to identify particular characteristics of the collaboration processes and data-sharing practices. Then the items were divided to distinguish between behavioural and attitudinal features. Table 3.2 shows the list of the items and its source from the literature review.

Table 3.2: Interview guide items. We categorise the interview guide’s item by the two themes of this inquiry and provide the sources that define each construct item.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Items</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaboration</td>
<td>compatibility</td>
<td>(Hara et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>decision making</td>
<td>(Thomson, Perry, and Miller 2008)</td>
</tr>
<tr>
<td></td>
<td>purpose</td>
<td>(Roberts and Bradley 1991)</td>
</tr>
<tr>
<td></td>
<td>organisation</td>
<td>(Roberts and Bradley 1991)</td>
</tr>
<tr>
<td></td>
<td>autonomy</td>
<td>(Harley and Blismas 2010)</td>
</tr>
<tr>
<td></td>
<td>complementarity</td>
<td>(D’Amour et al. 2005)</td>
</tr>
<tr>
<td></td>
<td>mutuality</td>
<td>(D’Amour et al. 2005)</td>
</tr>
<tr>
<td></td>
<td>trustworthiness</td>
<td>(Harley and Blismas 2010)</td>
</tr>
<tr>
<td></td>
<td>motivations</td>
<td>(Hara et al. 2003) (Ryan and Deci 2000)</td>
</tr>
<tr>
<td></td>
<td>infrastructure</td>
<td>(Hara et al. 2003) (Ryan and Deci 2000)</td>
</tr>
<tr>
<td></td>
<td>data sharing</td>
<td>(Harris and Miller 2011)</td>
</tr>
<tr>
<td></td>
<td>commoditization</td>
<td>(Ryan and Deci 2000)</td>
</tr>
<tr>
<td></td>
<td>motivations</td>
<td>(Harley and Blismas 2010)</td>
</tr>
</tbody>
</table>

Purpose includes elements such as responsibility and goal clarity and Organisation includes elements such as coordination and meeting efficiency.
Data Collection Procedure

We followed a simple data collection procedure. We contacted the subjects via email and through an introduction via the SDR development team. Then, we administrated the semi-structured interviews to the subjects who accepted the invitation. Each interview took between 30 minutes and 1 hour to finalise. We audio-recorded the interviews and then transcribed them. Finally, we sent a closed questions survey to confirm the information collected during the interviews.

We interviewed all the subjects from June to November 2013 and performed the analysis in May 2014. We interviewed some of the subjects at their locations (in Germany and the UK) and some others over the phone. We kept all the data in the main researcher personal computer in an RQDA\(^3\) encoded file. We used a verbatim way for interview transcription. We anonymised the interviews using four character codes (e.g. SI–01, SI–02, SI–n).

Participant Selection

<table>
<thead>
<tr>
<th>Item</th>
<th>Category</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>3</td>
</tr>
<tr>
<td>Location</td>
<td>Germany</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>United Kingdom</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Netherlands</td>
<td>2</td>
</tr>
<tr>
<td>Discipline</td>
<td>Modeller</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Experimentalist</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1</td>
</tr>
<tr>
<td>Career Stage</td>
<td>PhD student</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Post-doc</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Research Assistant</td>
<td>2</td>
</tr>
</tbody>
</table>

n: Number of participants

3.1. EXPLORATORY CASE STUDY

Figure 3.2: Interviews participation. Showing a 17% drop-off rate.
additional participants who meet the inclusion criteria and could potentially contribute to the study [Yin13]. The approach allowed us to contact participants that we would not have been able to reach otherwise.

We used an inclusion criterion with the following characteristics: (1) participants had to be active in the scientific collaboration, and (2) they must have used the SDR. Figure 3.2 presents the study participation. We asked 56 researchers to participate; 11 consented, and 10 had completed the interview. As shown in Table 3.3 the participants came from different geographical locations; we obtained a balanced sample of the two foremost disciplines, and all of them were in an early-career stage.

Repository Logging data

We used data-logs from the collaborations’ Scientific Data Repository to enrich the study’s observations. To make observations of the scientists’ data-sharing practices, we used two metrics. First, we used the proportion of restricted datasets to accessible datasets as a data-sharing proxy measure. This metric is a proxy because scientists could be sharing data in other repositories or archives to which access was not available. Similarly, it is possible that they were not depositing all the data they produce in the repository. The second measure we made use was the of some self-deposited datasets. Although these metrics did not represent the actual amount of produced or shared datasets, it did give us an indication of the scientist’s practices in the SDR.

3.1.5 Data Analysis

We assessed the interview data as follows. First, all interview transcripts were coded. For the coding of the interview transcripts, we used descriptive coding, which is characterised by summarising the primary topic of the excerpt in the interview transcript [Sal12]. Then, the codes were classified to identify categories and themes. Finally, we printed codes out onto cards, and we performed a reverse card sorting with a second coder to evaluate the categorisation [Gor+98]. We selected reverse card sorting as evaluation instrument because it provided the right balance between coding reliability and resources.

We used two metrics to analyse the SDR’s logs. We defined self-deposition using the average amount of competitive assets self-deposited over a funding
period. A competitive asset is that is not usually published (i.e. Data in spreadsheets, mathematical models and SOP’s). We analysed research-outputs using the average amount of articles published over a funding period. We analysed data-sharing using the average amount of articles published over a funding period. We standardised all metrics for their comparison between collaborations.

A list of data [Gar16e] and electronic analysis tools [Gar16a] used here can be found in Appendix B.

3.1.6 Results

Four broad themes emerged from the analysis. The first theme concerns the collaborations’ participation. The second is about the collaborations’ research process. Finally, the last two themes are about the collaboration’s data-sharing Practices and Attitudes. In the following pages, we describe these themes.

Invisible Small Collaborations

From a methodological perspective, we identified a trend with regarding the composition of the collaborations (cases) that we studied and those that did not reply to our invitation for study. Both sets of collaborations were under the same programme (i.e., Galaxy Groups Programme) hence we were able to collect some descriptive data about all of them. We found that the collaborations we analysed had three contrasting differences with those that we were not answered our invitation. These three characteristics were: levels of dataset self-deposition; levels of reported results; commitment with data management (Fnd1).

Table 3.4 shows the collaboration’s self-deposition per year scores and reported results per year scores. The collaborations with the lower scores were among those that did not respond to this study’s invitation for participation.

Moreover, in terms of commitment to data-sharing, the Data Management team indicated that this set of collaborations showed less commitment to the data management goals (Fnd2). Talking about this issue, one of the Principal Investigators (PIs) said:

*The projects STEPHAN and ROBERT had very good liaison partners, but their PIs never bought into the idea of data-sharing and data management...* – [PI–01]
Table 3.4: Comparison of reported collaboration’s contributions as of 2015. Showing the collaborations’ reported outputs and invitation response.

<table>
<thead>
<tr>
<th>Collaboration</th>
<th>Outputs</th>
<th>Self-Deposing</th>
</tr>
</thead>
<tbody>
<tr>
<td>WILD</td>
<td>1.17</td>
<td>-1.16</td>
</tr>
<tr>
<td>MARKARIAN</td>
<td>-0.95</td>
<td>-1.12</td>
</tr>
<tr>
<td>ROBERT</td>
<td>-0.21</td>
<td>-0.99</td>
</tr>
<tr>
<td>SEYFERT</td>
<td>0.77</td>
<td>-0.91</td>
</tr>
<tr>
<td>STEPHAN</td>
<td>-0.95</td>
<td>-0.71</td>
</tr>
<tr>
<td>BULLET</td>
<td>0.04</td>
<td>-0.16</td>
</tr>
<tr>
<td>BURBIDGE</td>
<td>-0.37</td>
<td>-0.13</td>
</tr>
<tr>
<td>ZWICKY</td>
<td>0.03</td>
<td>-0.13</td>
</tr>
<tr>
<td>COPELAND</td>
<td>-0.62</td>
<td>0.24</td>
</tr>
<tr>
<td>DEER-LICK</td>
<td>0.04</td>
<td>0.66</td>
</tr>
<tr>
<td>LEO</td>
<td>2.25</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Quantities are normalised for the comparison.
* Obtained a funding extension after the first year evaluation.
◦ Responded to our study’s invitation request.
n Joined the programme during the second round of funding.

This view was reflected by the fact that these collaborations had the lowest ratio of collaborators to contributors. Additionally, all these collaborations were not funded for a second round by the consortium.

The Research Collaboration Process

A recurrent theme in the interviews was a sense amongst interviewees that decision making in the small collaborations was characterised by general agreement and healthy discussions (Fnd3). As one interviewee said:

\[ \ldots \text{For the modelling we have always had different points of view with the [collaborators]. I had some healthy disagreements with some decisions, but I also understood we had to make some concessions. – [SI–10]} \]

One participant commented:

\[ \ldots \text{We always found some consensus. We had some hard discussions... But this is normal. I think we never left a consortium meeting without a real consensus about how to proceed. – [SI–02]} \]
Another interviewee, when asked about their collaboration’s decisions, said:

\[\ldots\text{ There were internal discussions among our group, and we work}\
\ldots\text{ closely with one of the other laboratories in [remote location].... We}\
\ldots\text{ sat together, and we discussed a decision to suit everyone} – [SI–04]\]

In all cases, participants reported a **broad awareness of their research goals and responsibilities** (**Fnd4**). By this, we mean that collaborators showed an awareness of their responsibilities, others’ responsibilities and their collaboration goals. Talking about this issue an interviewee said:

\[\ldots\text{ [the collaboration] had three groups doing experiments and three}\
\ldots\text{ groups doing modelling. My group did research on SLAB and [the other group] did the modelling.} – [SI–01]\]

On contrasting the responsibilities of the collaborators one subject stated that:

\[\ldots\text{ the other group investigated DNA damage. In the [remote institution], they studied perturbation of transporters... We handed out our experimental data and our results to those groups} – [SI–07]\]

Moreover, another commented:

\[I\text{ was responsible for the messenger RNA and transcription while my collaborator at [remote location] was responsible for measuring enzyme kinetics...} – [SI–04]\]

Another recurrent theme in the interviews was a sense amongst interviewees of coordination problems with remote collaborators (**Fnd5**). For example, one interviewee said:

\[\ldots\text{ Inside [my institution] we had a very good collaboration. However,}\
\ldots\text{ with members outside there was not such a close relationship or good}\
\ldots\text{ communication. I had a collaborator, an experimentalist, in [remote location] and I needed data from him, but it was always easier working with the experimentalist from [my institution]} – [SI–03]\]

One informant reported that:
I think it would have been better if the same group would have performed all the research and experiments in the same university... I was working alone at my institution and I think it would have been better if I had been physically at the university where the experiments were performed. – [SI–10]

Another interviewee alluded to the notion of coordination timing:

all collaborators (modellers and experimentalists) started working at the same time... [but] during the first year I had no data to provide to the other groups. This was a problem for them. I could not produce data for them from day one. It was a pity – [SI–06]

A common view amongst interviewees was that all small collaborations had shown an absence of autonomy conflict (Fnd6). By that, we mean that their autonomy was not compromised by the research goals of the collaboration. That was generally characterised by the working status of the collaboration members. Except for one case, all the interviewees reported that all their work was with the collaboration. Thus their goals and responsibilities were always aligned.

We found that collaborators expressed a variety of perspectives about the way small collaboration’s members complemented each other (Fnd7). In other words, they expressed that, the small collaborations needed each of the members for both resources, such as data and instrumentation, as well as access to expertise from other scientists. As one interviewee put it:

I could not have performed the research if I have not had been part of the collaboration with the experimentalist. [My] model by itself means nothing without the experimental setup [the other collaborators provided] – [SI–08]

Another informant echoed this view:

There is no way our group would have done this work without the collaboration. [without them] this type of data and this type of experiment would not be available to us. – [SI–02]

Talking about the issue of access to expertise an interviewee said:
3.1. EXPLORATORY CASE STUDY

I personally was learning about how modelling works and different types of modelling. These are people doing crazy stuff and to understand why they do it and how they do it, was be a great achievement.
– [SI–05]

We found that a common view amongst interviewees was that collaborators in each of the small collaborations were very trustworthy with one another (Fnd8). One participant commented:

[The collaboration] was very open. We were a small collaboration and everybody knew each other. I think we had a lot of trust in the small group. We share data-files with the whole collaboration – [SI–02]

Talking about this issue an interviewee said:

We had no problems sharing unpublished data with our collaborators. We followed a parallel route. We emailed the modelling groups the experimental data every time they requested. – [SI–07]

This view was echoed by another subject who reported:

I allowed access to all my unpublished work to the people I collaborated with because I trust them... – [SI–04]

In their accounts of the events about compatibility, we found that some interviewees argued that they experienced problems with different working styles while others experienced complications around the scientific approach of collaborators (Fnd9). Some participants expressed that:

At the beginning, it was really difficult to work together... We noticed that the mathematical modellers... had very high expectations, and they were very organised. They want everything nice and perfect, and we [the experimentalist] cannot do that because the data do not allow for that– [SI–01]

Another interviewee, an experimentalist, alluded to the notion of different scientific approaches:
Communication [with mathematical modellers] was complicated at the beginning. Sometimes, they would overestimate our data. Sometimes it was complicated, but we solved that problem with frequent communication. – [SI–06]

Another interviewee said:

we had a difficult stage at the beginning. We went to the lab and sat with the experimentalist trying to understand their scientific approach... But after that we did not have any problems. – [SI–03]

Overall, these results indicate three major takeaways. First, regarding the collaborative process, all collaborations were very similar. Additionally, collaborations were characterised by good levels of trust and commitment to the goals. Finally, the few problems they encountered were related to the heterogeneity of the collaborators discipline and the collaborator’s geographic distribution.

Data Sharing Practices

We found that collaborations followed different data-sharing practices were correlated to the data life cycle [UK 10] and academic publishing life cycle (Fnd10). We used the term “dataset” to refer to any scientific material produced by the scientific collaborators (e.g. spreadsheets, mathematical models, SOPs, etc.).

Our findings indicate that pre-article-publication data-sharing only occurs between collaborators (i.e., Specific Sharing) and only in some cases relies on the use of the data repository. Talking about this issue, an interviewee said:

[If I share my data] without restrictions people do not know whether it is published or not published. That is an issue. What if they use it? I shared [my data] with collaborators because I know them. I trust them. So, nobody is working with my data without asking me first. – [SI–04]

Talking about this issue an interviewee said:

I shared data only in a closed environment... Although I am the expert within the project, other groups in the world could get a competitive advantage if they get access to it – [SI–10]
3.1. EXPLORATORY CASE STUDY

Interviewees also suggested that:

We normally gave access to members in our collaboration... Sometimes we made temporary links available when we had manuscripts under [peer-review] – [SI-06]

These views were supported when observing the access status of datasets in the SDR years after their deposition. We considered all datasets in the repository in their initial deposition. We found that the majority of these datasets remained restricted to public access years after their first deposition. This observation is reflected in [Figure 3.3] which shows the percentage of data each collaboration shared with collaborators in the repository (i.e., Specific Sharing) to the percentage of data shared openly (i.e., Data Publishing).

We found a disparity between the scientists’ perception of data publishing events and the manifestation of data publishing events (Fnd11). The substantial majority of the participants expressed that data was being shared after article publication. For example, one interviewee said:

[We did not share data] because the collaborators were waiting to write a paper by the [experimentalists] and publish it... It is just a matter of publication. First, they get a publication then we do [the data] sharing. – [SI-03]

One informant reported that:

... once any publication would get released we would make the datasets fully open to the public... – [SI-06]

Another informant echoed this view:

... It was only when we were sure that our results were acceptable then we made the datasets open access. That was generally after publishing [an article]... – [SI-07]

This claim was contrasted with evidence currently existing on the data accessible on the repository: Many datasets remained restricted even after article publication. We were not able to map dataset to reported-publications (i.e., articles). However, we observed that a large percentage of dataset remained under
Figure 3.3: Data Published and data shared with collaborators (i.e., Specific Sharing) from the six collaborations. Most of the collaboration kept their data hidden away from the public and published less that 25%.

restricted access (as shown Figure 3.3) even after years of the last-reported article being published. Table 3.5 shows the last activity performed in the repository by the collaborations, as late as 2015, against the date of the last article publication from that collaboration.

Additionally, we found that the commitment to the data-sharing agreements was not present. One informant reported that:

*with all honestly we do not mention [the repository] anymore, not even the mention of data-sharing... I am sure our bosses do not even log in to [the repository] to see what’s happened.* – [SI-04]
3.1. EXPLORATORY CASE STUDY

Table 3.5: Comparison between last publication and last activity in the SDR. Showing that all collaborations have not been active in the SDR after publishing manuscripts.

<table>
<thead>
<tr>
<th>Collaboration</th>
<th>Last Activity</th>
<th>Most Recent Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copeland</td>
<td>2013-01-06</td>
<td>2013-05-03</td>
</tr>
<tr>
<td>Leo</td>
<td>2013-01-07</td>
<td>2013-12-05</td>
</tr>
<tr>
<td>Bullet</td>
<td>2013-02-12</td>
<td>2014-09-30</td>
</tr>
<tr>
<td>Zwicky</td>
<td>2013-04-03</td>
<td>2014-12-01</td>
</tr>
<tr>
<td>Burbidge</td>
<td>2013-01-30</td>
<td>2014-12-19</td>
</tr>
<tr>
<td>Deer-Lick</td>
<td>2013-01-09</td>
<td>2015-08-01</td>
</tr>
</tbody>
</table>

Talking about the assets they kept private, an interviewee said:

*I totally just forgot to share [some of my published assets]. – [SI–08]*

It was also suggested that commitment towards sharing was low:

*In our first year, our “liaison” member was putting pressure after realising that nobody was uploading anything [to the repository]. I think that probably happened with most of the groups during the first year. – [SI–06]*

We found that mathematical modellers perceived that Experimentalists have unique data-sharing practices (*Fnd12*). A small number of those mathematical modellers who were interviewed suggested that:

*Modellers would upload preliminary datasets [to the repository]. However, Experimentalists would only upload final datasets –[SI–05]*

Talking about this issue a modeller said:

*I would have no problem with sharing the dataset but [Experimentalist] are first concerned with publishing an article... Often I would have to wait until they published to make my dataset open access –[SI–03]*

Another Modeller commented:

*As a mathematical modeller I would like to make my research open, but most experimentalists tried to keep their resources under control –[SI–08]*
Data Sharing Attitudes

We found that collaborators’ data-sharing motivations come purely from external sources. In other words, their motivations were extrinsic \( \text{RD00} \) (Fnd13). As one interviewee said:

\[
\ldots \text{[depositing data in the repository] was done to make sure that we as a collaboration would get our second funding. Because our bet was that [way] we would be able to continue [getting funding].} - [SI–01]
\]

Some participants expressed the belief that:

\[
\text{In general collaborators would put data in [the repository] at the end of every review year. When we knew people would check our profile. That was the usual case in our collaboration.} - [SI–08]
\]

This view was echoed by another interviewee who reported:

\[
\text{[Self-depositing in the repository] was a part of the legal obligation of the project funding. We were just respecting that.} - [SI–04]
\]

Moreover, this attitude was reflected in practice. We considered funding evaluation datelines as an extrinsic incentive for dataset self-depositions. As the interviewees expressed, the funders and bosses would examine the repository at these dates to evaluate whether the collaborations were worthy of further funding. Figure 3.4 shows that around every funding evaluation date, datasets self-depositions to the SDR consistently spiked. This trend supports the perception of the interviewees.

On the whole, we found that there was a perception that data-sharing provided no individual benefits (Fnd14). As one interviewee put it:

\[
\text{I did not think uploading my data would affect my career, because of visibility, because [the repository] is less visible.} - [SI–04]
\]

This view was echoed by another interviewee who said:

\[
\text{specifically for myself, I do not see how having shared my data would be beneficial. However, I can understand its beneficial for others.} - [SI–10]
\]
Figure 3.4: Distribution of self-depositions over time by the six collaborations. Showing a relationship between rapid increases of self-deposition and funding evaluation deadlines.

It was also suggested that:

\(\ldots\) most scientists [share their data] only because they have the feeling the committee would be looking at them. – [SI–01]

Additionally, we found that fear of getting scooped was among the main reasons behind data withholding (Fnd\textbf{15}). One informant reported that:

In my case, I do not see much of a problem with data-sharing but other collaborators are working on topics that are more competitive, and they would like to keep some stuff secret from their competitors. – [SI–06]
This view was echoed by another informant who said:

_There was trust with the collaborators. However, there was mistrust about putting the data on the Internet. People have a right to keep their data._ – [SI–05]

And:

_I shared data only in a closed environment... Although I am the expert within the project, other groups in the world could get a competitive advantage if they get access to it._ – [SI–10]

We found that participants disagreed on the objectives of the SDR. The overall consensus was that the SDR was used to publish data (Fnd16). This attitude was present in practice; see section 3.1.6. Moreover, this was reflected in the collaboration’s data-sharing Policy in June 2011 [Pas11]. On the other hand, there were some suggestions that collaborators saw the SDR as a tool to help the challenges of remote collaborating. Some participants expressed the belief that:

... as a mathematical modeller, I want to work with the preliminary data. I do not want to wait until an [experimental] dataset is completely finished and checked... But the collaboration decided to use the repository only for data that are complete and evaluated. Making the whole data repository a little bit obsolete for data exchange between the collaboration partners – [SI–02]

There were some suggestions that all collaborations were using the SDR as a collaborative tool as well as a publishing tool. Some participants expressed:

...we normally shared our assets with all our collaboration in the repository before publishing an article. – [SI–06]

Another interviewee said:

... We deposited datasets in the repository to exchange them between collaborators. Sharing between each other is the repository’s beneficial feature. – [SI–04]

A summary of the findings identified in this study is presented in Table 3.6. In the next section we compare and contrast our findings with the literature in order to discuss them.
3.1.7 Exploratory Study Discussion

Remote Collaborations

One unanticipated finding was that established collaborations with low levels of reported outcomes and dataset depositions were less prompt to engage with this study's invitation for participation (Fnd1). There are three potential explanations for this result. A probable explanation for this might be that members from these collaborations are less prompt to interact because they considered the collaboration unsuccessful. Another potential explanation might be the lack of leadership on the topic of data-sharing. One of the PIs suggested this. Stokols et al. [Sto+08] suggests that leaders in trans-disciplinary collaborations influence collaborative processes and outcomes. It may be that these collaborations had no interest in participating because they stopped working together years before the participating collaborations.

On the other hand, we found that the studied collaborations had shown a number of positive aspects regarding collaboration process (findings Fnd2..Fnd9). Specifically, responsibility awareness and the decision-making process. These results could be related to the fact that the collaborations were defined before the working members joined. Responsibilities were already defined in roles and goals were clear from the outset for all participants. In his study of the process of cross-disciplinary research collaboration, Jeffrey [Jef03] suggests that, to have a smooth collaborative process, responsibility for the products of collaboration should be clearly assigned from the beginning. The latter has been echoed by Bammer, who has shown that, by explicitly planning for and managing interdisciplinary interactions, a high degree of data, information and knowledge synthesis can be achieved in a scientific collaboration [Bam08]. Another potential reason behind this alignment between collaborators could be the consensus on the decision-making process. Tapia et al. [Tap+11] suggested that a lack of early consensus in inter-institutional collaboration quickly becomes a barrier to Collaboration Readiness.

The two problems expressed by the interviewees were related to the remote aspect of the collaborations. This issue with remote collaborators has been pointed out in the literature (e.g. [Sto+08], [WM07], [Tap+11]). In their survey of scientists in experimental biology, mathematics, physics, and sociology, Walsh and Maloney [WM07] found that remote collaborations are particularly prone
<table>
<thead>
<tr>
<th>Table 3.6: Exploratory Study Findings.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Categories</strong></td>
</tr>
<tr>
<td><strong>Finding</strong></td>
</tr>
<tr>
<td><strong>Theory</strong></td>
</tr>
<tr>
<td><strong>Practices</strong></td>
</tr>
<tr>
<td><strong>Attitudes</strong></td>
</tr>
<tr>
<td><strong>Findings</strong></td>
</tr>
<tr>
<td><strong>Fnd1</strong> Unproductive collaborations were less supportive.</td>
</tr>
<tr>
<td><strong>Fnd2</strong> Low collaboration commitment translated into low data sharing commitment.</td>
</tr>
<tr>
<td><strong>Fnd3</strong> Shown good decision making characteristics.</td>
</tr>
<tr>
<td><strong>Fnd4</strong> Collaborators have shown a broad awareness of their goals and responsibilities.</td>
</tr>
<tr>
<td><strong>Fnd5</strong> Collaboration initially struggled to work remotely</td>
</tr>
<tr>
<td><strong>Fnd6</strong> No threats to autonomy.</td>
</tr>
<tr>
<td><strong>Fnd7</strong> Presence of resource and expertise compatibility</td>
</tr>
<tr>
<td><strong>Fnd8</strong> Collaborators trust each other.</td>
</tr>
<tr>
<td><strong>Fnd9</strong> Collaboration initially struggled to get to know each other.</td>
</tr>
<tr>
<td><strong>Fnd10</strong> Collaboration initially struggled to work remotely</td>
</tr>
<tr>
<td><strong>Fnd11</strong> Misalignment of individual perception</td>
</tr>
<tr>
<td><strong>Fnd13</strong> “Sticks” only improve self-deposition</td>
</tr>
<tr>
<td><strong>Fnd14</strong> Lack of perceived benefits</td>
</tr>
<tr>
<td><strong>Fnd16</strong> SDR’s Objectives Misalignment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Findings</strong></th>
<th><strong>Findings</strong></th>
<th><strong>Findings</strong></th>
<th><strong>Findings</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Working styles</strong></td>
<td><strong>Working styles</strong></td>
<td><strong>Working styles</strong></td>
<td><strong>Working styles</strong></td>
</tr>
<tr>
<td><strong>Fndg</strong> Collaborations initially struggle to work remotely</td>
<td><strong>Fndg</strong> Collaborations initially struggle to work remotely</td>
<td><strong>Fndg</strong> Collaborations initially struggle to work remotely</td>
<td><strong>Fndg</strong> Collaborations initially struggle to work remotely</td>
</tr>
<tr>
<td><strong>Fndh</strong> Collaborations trust each other.</td>
<td><strong>Fndh</strong> Collaborations trust each other.</td>
<td><strong>Fndh</strong> Collaborations trust each other.</td>
<td><strong>Fndh</strong> Collaborations trust each other.</td>
</tr>
<tr>
<td><strong>Fndi</strong> Presence of resource and expertise compatibility</td>
<td><strong>Fndi</strong> Presence of resource and expertise compatibility</td>
<td><strong>Fndi</strong> Presence of resource and expertise compatibility</td>
<td><strong>Fndi</strong> Presence of resource and expertise compatibility</td>
</tr>
<tr>
<td><strong>Fndj</strong> No threats to autonomy.</td>
<td><strong>Fndj</strong> No threats to autonomy.</td>
<td><strong>Fndj</strong> No threats to autonomy.</td>
<td><strong>Fndj</strong> No threats to autonomy.</td>
</tr>
<tr>
<td><strong>Fndk</strong> Their goals and responsibilities are clear.</td>
<td><strong>Fndk</strong> Their goals and responsibilities are clear.</td>
<td><strong>Fndk</strong> Their goals and responsibilities are clear.</td>
<td><strong>Fndk</strong> Their goals and responsibilities are clear.</td>
</tr>
<tr>
<td><strong>Fndl</strong> Show data sharing commitments</td>
<td><strong>Fndl</strong> Show data sharing commitments</td>
<td><strong>Fndl</strong> Show data sharing commitments</td>
<td><strong>Fndl</strong> Show data sharing commitments</td>
</tr>
<tr>
<td><strong>Fndl</strong> Low collaboration commitment translated into low trust</td>
<td><strong>Fndl</strong> Low collaboration commitment translated into low trust</td>
<td><strong>Fndl</strong> Low collaboration commitment translated into low trust</td>
<td><strong>Fndl</strong> Low collaboration commitment translated into low trust</td>
</tr>
<tr>
<td><strong>Fndm</strong> Unproductive collaborations were less supportive</td>
<td><strong>Fndm</strong> Unproductive collaborations were less supportive</td>
<td><strong>Fndm</strong> Unproductive collaborations were less supportive</td>
<td><strong>Fndm</strong> Unproductive collaborations were less supportive</td>
</tr>
</tbody>
</table>

**Challenges**

- **Fnd1** Fear of scooping
- **Fnd2** Lack of perceived benefits
- **Fnd3** "Sticks" only improve self-deposition
- **Fnd4** Misalignment of individual perception
- **Fnd5** Misalignment of disciplinary perception
- **Fnd6** "Sticks" only improve self-deposition
- **Fnd7** Low collaboration commitment translated into low data sharing commitment
3.1. EXPLORATORY CASE STUDY

To problems of coordination and misunderstandings. Furthermore, Stokols et al. [Sto+08] suggested that remote collaborators must adapt to the loss of shared physical settings and socio-spatial cues in order to make the collaboration work.

We found that collaborators also encountered conflicts with different working styles and research approaches. The conflict might be related to the discipline and institutional diversity in the collaborations. For instance, Jeffrey’s study of cross-disciplinary collaborations between modellers and sociologists [Jef03] points out that, in order to function effectively, different skills are required when working at a cross-disciplinary research collaboration to thus needed when working within a small disciplinary team. Concerning institutional diversity, Heinz and Kehlmann’s study of the German public research system [HK08] suggested that incompatible working routines can hamper inter-institutional collaborations.

The good degree of trust we encountered in the collaborations could be explained by initial face-to-face contact and constant socialisation. Stokols et al. [Sto+08] have previously suggested that, the latter factor, might be right. They argue that initial face-to-face contact and appropriate communication can increase the trust levels among team members, facilitate the formation of social norms, and aid the establishment of group identity.

Practices

The current study found that collaborators followed different data-sharing practices accordingly to when they found the times to be less risky (Fnd10). Collaborators expressed that data-sharing would solely be performed when their datasets were “protected” by a publication endorsement. One mechanism that could explain this is Loss Aversion. Loss aversion is people’s tendency to prefer avoiding losses strongly as opposed to acquiring gains [TS08]. In this case, collaborators expressed their fear of losing control over the use of their data and their fear of getting no attribution for their reuse.

These results are in agreement with those obtained by some studies. For example, in their case study of a Systems Biology community, Sheridan et al. [RIN08] found that some scientists prefer to keep the datasets private for a period until they have exploited the dataset to the best of their knowledge. This is not only restricted to Systems Biology. In his case study of five disciplines from the Physical Sciences Cragin [Cra+10], found that more than half of their subjects identified a need to restrict some, or all, of their dataset from public...
access for some length of time. Additionally, similar results were found by the
survey conducted by Tenopir [Ten+11a] in which subjects expressed the re-
quirement of conditions on access to the data they deposit in a data repository.
Furthermore, recent studies, that were not available at the time we performed this
research, also supported these results (e.g. Jessel and Birch Survey of early users
of Figshare [JB14] and Youngseek and Stanton multilevel analysis of researchers
from Biological Sciences [KS15]).

Another significant finding was that a disparity between the scientists percep-
tion of Data Publishing events and the manifestation of these events (Fnd11).
The majority of the participants expressed that data was being shared after ar-
ticle publication and showed surprise when they realised some of their data was
still behind the wall. A possible explanation for this might be that participants
were not committed to goals and thus did not pay attention to the outcomes of
the repository. Another possible explanation for this is that leadership was not
aligned with the data-sharing policy. Burke et al. [Bur+06], outcomes in their
meta-analysis of leadership, have highlighted the role of leadership when aiming
to achieve collaboration performance. These results are aligned with those of
Savage [SV09]. In his study of datasets’ requests from clinical trials, he found
that, even when explicit data-sharing agreements were established, scientists did
not share their data upon request. The results are also consistent with the lack
of data-sharing commitment that we found in the subjects (Fnd16).

We found that mathematical modellers perceived that Experimentalists have
different data-sharing practices (Fnd12). A possible explanation for these results
may be the differences in practices between secondary data users and primary
data users. Another possible explanation for this is the discipline’s degree of
technology acceptance. Technology Acceptance is an information systems theory
that models how users come to accept and use a technology. It states “the degree
to which a person believes that using a particular system would enhance his or
her job performance” [Bro03, Hei04]. It could be that experimentalists, unlike
modellers, do not perceive their work to be enhanced by using the repository or
for sharing data. Another possible explanation for this is the disciplines’ stage
in the innovation diffusion process. The innovation diffusion process states that,
when an innovation is introduced, its diffusion happens through different groups
of members in a sequential fashion. The first group of people to use a new
technology are called “innovators,” followed by “early adopters.” Next come the
early and late majority, and the last group to eventually adopt the innovation are called “laggards” [Rog10]. It is possible that the modellers are one or two stages ahead of the experimentalist on this process and just time would help them to get to the same stage. These results are in line with those of previous studies on data-sharing practices. For example, different self-reported studies have given accounts of differences in data-sharing behaviour between disciplines (e.g. [RIN08], [Cra+10], [Ten+11a]). Similarly, analyses of scientific data repositories’ deposition rates have shown that discrepancies in the way disciplines deposit data exist (e.g. [Piw11] [PCI10]).

Attitudes

We found that, although attaching rigid norms to extrinsic incentives improves data self-deposition, it does not improve data-sharing (Fnd13). A possible explanation for this is that collaborators might have realised that their reputation would not be damaged when they did not perform data publishing. It could be argued that there are two types of stakeholders in these collaborations: the collaborators and the funders. The collaborators knew that the lack of datasets deposition on the SDR would reduce their chances of getting the approval of the funders for another funding period (e.g. the three projects with the lowest deposition scores were not given funding for the second period). Moreover, collaborators knew that, by sharing data internally, they would have a good reputation among collaborators. By meeting these two conditions, they kept their reputation in good shape with both stakeholders. One could argue that the collaborator’s reputation would be damaged when they failed to conduct Specific Sharing, as the data-sharing policy explicitly required. However, the data-sharing policy was only enforceable during the length of the project. Publishing data was only enforced after article publication, which, in some cases, can take months or years beyond the end of the collaboration project. Therefore, collaborators could have always been excused from not sharing their data publicly by using the article-publication-card. These results are in agreement with Savage’s [SV09] findings, which showed that explicit data-sharing policies in journals do not lead authors to share data.

Even when self-deposition relies mostly on extrinsic incentives (e.g. funding cuts), subjects presented a narrow band knowledge of other benefits that could act as extrinsic incentives (Fnd14). The reason for this is not clear, but it may
have something to do with lack of awareness of individual benefits. If they have not experienced the individual benefits of data-sharing, they are less likely to be aware of them. Most of the researchers were early academics so they are unlikely to have experienced benefits such as creating new collaborations as a result of data-sharing [KH09]. In contrast, previous studies of other cohorts have shown a broader awareness of individual benefits. In their study in the field of Biodiversity, Enke et al. [Enk+12] found that one of their main motivations to share data is the availability of comparable data sets for comprehensive analyses. Youngseek and Stanton’s survey [KS12] identified that researchers expressed a belief that data-sharing could highlight the quality of their work in research. Additionally, the PARSE-Insight Survey [KH09] identified that researchers believe that data-sharing may stimulate interdisciplinary collaborations.

We also found that fear of being scooped ranked high in the reasons for data withholding even when they had not experienced scooping in the past (Fnd15). This attitude could be connected with the collaborators’ perception of risk. Collaborators might think there are no safeguards for data reuse attribution other than morality. A recent study by Mooney et al. [MN12] has highlighted that citation of data is not yet a norm in many fields of science. Outcomes citation is the primary retribution mechanism in academia. By not enforcing the data citation there is no guaranteed of retribution on data reuse. These results match those observed in earlier studies. For example, Campbell’s survey in the field of Genetics [Cam+02] found that researchers are mostly concerned with the way in which data-sharing could jeopardise their junior faculty members’ ability to publish. Cragin et al. [Cra+10] found similar fears in researchers from across different fields (i.e. Agronomy, anthropology, earth and atmospheric sciences, geology, and plant sciences). Similar concerns have been raised by Bioinformaticians in the survey conducted by Haeussler [Hae+13] and by researchers in Synthetic Chemistry in the case study conducted by Velden [Vel13]. Furthermore, more recent studies, which were not available at the time we performed this study, also supported these results (e.g. Jessel and Birch Survey of early users of Figshare [JB14] and Tenopir’s follow-up survey [Ten+15]).

Limitations

Although this research generates valuable insights for, both the specific case, the Galaxy Groups Programme and the small scientific collaborations in general,
we made a number of choices that impose some limitations to this study. In terms of case selection we have not had access to the full range of collaborative practice in the Galaxy Groups Programme. We were only able to collect data from successful collaborations. Consequently, this study is inevitably limited to the context of these collaborations. Therefore, we could only characterise collaborations that have some level of data-sharing practice and have engaged in collaborative work. As far as generalising these results with other small scientific collaborations, we were only able to collect data from successful collaborations; hence we have not had access to the full range of practice in small scientific collaborations. Therefore, the study is inevitably limited to the context of this collaboration type of arrangement and, to a certain extent, to the discipline of the study cohort, Systems Biology. Furthermore, in both instances, we have chosen to use data self-deposition (i.e., Specific Sharing) as a proxy measure of intent for data-sharing (i.e., Data Publishing). Consequently, these results can only claim a measure of intent, not actual data-sharing behaviour.

**Implications**

The findings of this study suggest that there are a number of implications for scholarly communication researchers, scientific collaboration researchers and SDR designers.

These findings implicate that small scientific collaboration researchers will need to pay particular attention to the formation of small collaborations. In order to identify the factors that will lead them to fail or succeed.

There are three important implications for scholarly communications researchers. In terms of the perception of individual behaviour one important implication of these results is that, researchers should be doubtful on using results from self-reported surveys to measure data-sharing. In these small collaborations, participants tended to overestimate the openness of their data-sharing practices. Concerning the perception of data-sharing practices by different disciplines these findings have raised important questions about the nature of these perceptions. We did not find evidence of disciplinary differences on datasets shared on the data repository but the interviewees’ perception was that this difference existed. Is it possible that the cooperators (i.e., modellers) adapted to the defectors (i.e., wet lab experimentalist) behaviour? Or this difference never existed?. Another implication of these findings is that, practitioners seeking to improve data-sharing
behaviour in small collaborations should target into reducing scientists perceived risk.

These findings suggest several courses of action for SDR designers. For instance, developing targeted interventions aimed at improving data-sharing should be incorporated to the researchers’ workflows. Additionally, there is a definite need for creating systems that would reduce the perception of risk or increase the perception of benefits when scientists share data.

Conclusions

We found that, even in effective small collaborations, data-sharing still suffers from some issues that have been present in the community of life sciences at large. We identified the different practices followed by the small collaborations and found that the execution of data-sharing under all these practices change in function of the risk associated with data-sharing. As was mentioned in the previous section, we identified that the collaborations are purely extrinsically motivated to execute data-sharing but only to the point to perform specific sharing (i.e., depositing in an SDR under restricted access). In summary, many of the factors that affect data-sharing in the Life Sciences community stand as factors influencing small collaborations within this cohort.

In the next section, we present alternative methodologies considered and piloted in order to study small collaborations sharing data. We then, present the application of the second step to capture context in small collaborations using an SDR: identifying discipline-specific factors associated with data-sharing in the literature.

3.1.8 Methodological Rationale

Besides the exploratory case study, a number of other approaches were considered and piloted to answer the research questions posed in this chapter. These alternative studies include a literature review, a longitudinal study, a survey and a multiple case study. Additionally, we attempted to measure data sharing using text mining in publications. This section describes these studies and their limitations in order to answer the research questions examined in this chapter and the problems we encountered while piloting them. Figure 3.5 and Figure 3.6 show an overview of this discussion.
The first contemplated study was a literature review on data sharing by small collaborations in the field of Systems Biology. However in the literature we encountered a limitation on the number \( n=2 \), \([\text{RIN08}] \) and \([\text{Har12}] \) of studies involving the independent variables (i.e., small collaborations in Systems Biology) and dependent variables under study (i.e., Data Sharing). Therefore, we proceed to contemplate other studies.

The second study that we reflected on conducting was a longitudinal study of a small collaboration from Systems Biology. However, in the longitudinal study, we faced problems related to feasibility. First, the considered case was reaching the end of its lifespan, and we would potentially not be able to measure the
dependent variable more than once. Second, participant retention is always a concern in this type of longitudinal study, and small scientific collaborations at the end of their funding period were considered to be a risky aspect in this case.

Third, we reviewed and piloted a survey with members of a small collaboration from Systems Biology. The advantage of the survey study was that we could reach a large number of members from the collaborations. However, on conducting the survey, we encountered participation problems. The survey’s invitation response rate was 12%. Although we did not use the survey’s results to evaluate our approach, we did use the collected data to inform our exploratory findings.

Another approach considered was conducting a multiple case study. The multiple case study design involved the selection of two or more cases which ought to have similar characteristics [Yin13]. On piloting this study, in addition to the Galaxy Collaborations, we studied the Virtual Liver Network collaboration (VLN). The VLN Collaboration is made up of seventy Systems Biology research groups distributed across Germany. The groups work as small collaborations. We collected data through interviews with 10 participants in Germany. We followed the data collection approach described in section 3.1.4. However, the VLN Collaboration Data Management Group was unable to provide data logs for this study. Data logs are the only type of information that we could use to investigate the collaborators’ behaviour. Therefore, the analysis of the study was not conducted further. Unlike the survey, the interviews collected were not further analysed and used in the evaluation of this research. Nevertheless, these data can be used retrospectively to analyse this collaboration in the future.

Concerning measuring data sharing, we aimed to identify an objective measure of data sharing. Previous studies have been focused on using attitudinal metrics. For example, Savage and Vickers [SV09] use the attitude of scientists towards sharing upon request and Tenopir et al. [Ten+11b] use attitude towards sharing in a repository. However, these metrics do not reflect scientists’ actual behaviour. We piloted a text mining approach to measure data sharing from a defined collaboration of scientists. The metric we proposed tracked publications’ metadata with SDR self-depositions using metadata in the collaboration’s repository with terms identified in the publication using the TerMine system from NaCTeM. However, the preliminary results of the metric shown a pair of scalability problems. In essence, we found that authors did not follow uniform data citing practices. That led to the presence of a large number of false negative
results in the majority of the sample. Therefore, we opted to use metrics that look at data sharing intention, such as collaborative sharing and data publishing.

3.2. REVIEW OF DATA SHARING ASSOCIATIONS

In this section, we apply the second step of this thesis’s approach to context-capturing; identifying factors associated with data-sharing behaviour from the literature. In order to determine the factors, we performed a synthesis of the literature on data-sharing within the literature of life sciences. We first introduce the steps that we took to create the synthesis. Then we present the results of the synthesis. The section then is concluded by reflecting on the approaches taken and the results.

We performed a literature review to identify factors associated with helping
or hindering data-sharing. Data-sharing has been studied using different approaches. A number of studies focused on attitudinal studies concerning the conditions under which scientists would share data (e.g. [Enk+12], [Ten+11a], [JB14] and [Ten+15]). Others studies have focused on theories that explain how data-sharing could potentially emerge (e.g. [BB03] and [Bor11]). However, limited efforts have been conducted on devising a comprehensive list of factors and conditions that influence data-sharing behaviour. Moreover, with some exceptions (e.g. [KFH12], study of bioinformatics, biomedical engineering, biomedical chemistry, health behaviour, and materials science) most of the research on factors associated with data-sharing behaviour has been carried out in cases of a single discipline (e.g. Planetary Sciences [VD11], Medicine [Piw11], Psychology [WBM11]). Therefore, in this review, we aim for identifying such factors, and the results can be considered to be a survey contribution (contribution C1 from section 1.4: A set of factors associated with data-sharing behaviour.).

3.2.1 Literature Search and Appraisal

We performed a search of studies that have tested or observed a correlation between an intervening factor and data-sharing behaviour. We chose three constraints to narrow down this review’s search; time frame, studied discipline and existence of empirical evidence.

1. A time frame. We included studies that were published after 2007. We selected this cut-off date based on the proliferation of data repositories and the introduction of new worldwide data-sharing policies and mandates.

2. Discipline. We included studies within a Life Sciences cohort in order to eliminate potential confounding factors related to the field of science. The literature reports a substantial variation in academic data-sharing across disciplinary practices. For instance, a recent multilevel analysis by Kim and Stanton [KS15] found that scientists’ data-sharing behaviour could be influenced by the type of scientific discipline. Similarly, Tenopir’s survey [Ten+11a] found significant differences between disciplines in scientists’ data-sharing management practices and perceptions. Therefore, limiting the study of life science literature, allowed us to exclude potential discipline-related confounding factors.
3.2. REVIEW OF DATA SHARING ASSOCIATIONS

3. Empirical Support. We included studies that have provided empirical evidence for their concluding correlation between data-sharing and intervention. Therefore, attitudinal studies and best practices studies were excluded from this study. This restriction will restrain the types of studies and consequently will provide us with a corpus of claims backed with strong evidence.

3.2.2 Synthesis of Factors

Using the defined strategy, we identified seven publications and thirteen factors about data-sharing that composed the initial corpus of claims in this synthesis. In one out of the thirteen factors, we found conflicting results.

Benefits, Costs, and Risks

Numerous surveys have explored how scientists perceive the associated cost, benefit and risk to data-sharing (e.g. [Kui09], [Ten+11a], [Enk+12], [JB14]). However, thus far, few studies have examined the association between the actual data-sharing behaviour and this perception. Kim and Stanton found, in two different studies (i.e., [KS12] and [KS15]) that perceived career benefit was found to have a significant positive influence on scientists’ data-sharing behaviour. In the same study, they found that the perceived career risk involved in data-sharing negatively influences a scientist’s data-sharing behaviour. Also, the perceived effort required to share data negatively influences a scientist’s data-sharing behaviour.

Mandates

The impact of mandates on data-sharing behaviour has recently gained traction. Surveys such as that conducted by [KS15] have shown that regulative pressure by funding agencies positively influences a scientist’s data-sharing behaviour. Piwowar [Piw11] analysed the bibliographic data from 11,603 publications and found that NIH funding levels were associated with increased prevalence of data-sharing. However, the prevalence of data-sharing behaviour remains at low levels. Piwowar [Piw11] warns that, even when funding levels were associated with increased prevalence of data-sharing, the overall probability of sharing remains low. Similarly, previous studies (such as [SV09] and [MP13]) have shown the
limitations of policy has had on improving data-sharing behaviour. Therefore, alternatives strategies to address data-sharing ought to be pursued.

Practices

Certain practices, such as article publication, the reporting of strong results, data production, and selecting embargoes, have been shown to be influential in data-sharing behaviour. In his study of data self-deposition in repositories, Kervin et al. [KFH12] observed that, when data-sharing did occur, it was a side effect of article publication. In her ethnographic study of collaborative data production, Vertesi [VD11], identifies that the context of data production determines the collaborative sharing. That is, collaborations which produced data in a genuinely collaborative fashion were more likely to share their data, rather than those that produced data in a competitive manner. In his analysis of data requests to published articles, Wicherts [WBM11] found an association between data withholding and the strength of the analysis in publications. That is, scientists who provided strong evidence in their published manuscripts were more likely to share their data upon request. Observations conducted by Roche et al. [Roc+14] on the Dryad repository, indicated that, when given the opportunity, scientists choose short embargo periods for the data related to articles which they have published.

External factors

Together with mandates and certain practices, such as publishing, some external factors have been found to be influential in prompting data-sharing. For instance, Kervin et al. study on Genomics [KFH12] showed that, within Genomics, voluntary data-sharing occurred only when a global crisis (i.e., influenza virus) was affecting the discipline.

Demographic characteristics

The literature has exposed relationships between a number of demographic characteristics of the population and data-sharing behaviour. These characteristics include discipline, attitude, age and gender.
For instance, in a case study of a multidisciplinary consortium of collaborations sharing data in an SDR, Harper et al. \cite{Har12} observed that wet-lab researchers were a “bottleneck” in the collaborations’ data-sharing activities. The researchers would delay the deposition of all other collaborators (from other disciplines) until they felt no risk of being scooped. Harper et al. \cite{Har12} work is complemented by Kim and Stanton \cite{KS15} multi-level analysis in the STEM fields. He found that embedded normative pressure in scientific disciplines positively influenced a scientist’s data-sharing behaviour.

Besides the normative pressure, the personal attitude of the researcher is also known to be a major factor that influences data-sharing behaviour. Surveys such as that conducted by Fecher et al. \cite{Fec+15} and Kim and Stanton \cite{KS15} have shown that researchers who highly value open access are significantly more likely to make their data available.

Surprisingly, studies that look at age and gender have provided some interesting findings. Piwowar analysis suggested that researchers with many years of experience are less likely to share data compared to those with fewer years of experience \cite{Piw11}. Contrastingly, Fecher et al. \cite{Fec+15} found that experience in academic research has no significant influence on the willingness to make data available. Regarding gender, Fecher found that female researchers were significantly less willing to make data available to other researchers than were their male colleagues \cite{Fec+15}.

### 3.2.3 Factors Analysis

Factors associated with data-sharing may be classified according to their manipulability into “dependent” and “independent”, as shown in Table 3.7. The term “dependent” refers to factors that can be modified by the SDR designers. In contrast, the term “independent” refers to factors that cannot be modified by designers. In the following, we exemplify the differences between these two types of factors.

**Independent factors**

Independent factors include factors that are characteristics of the users, scientific practices, mandates, or variables that are out of the direct control of the SDR designer. These factors must be considered with a note of caution because they
CHAPTER 3. TRANSLATING THEORIES

Table 3.7: Categorisation of factors associated with the data-sharing in the literature.

<table>
<thead>
<tr>
<th>Category</th>
<th>Factor</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent</td>
<td>Perceive Benefits, Costs and Risks</td>
<td>(Youngseek and Stanton 2015)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent</td>
<td>Mandates</td>
<td>(Youngseek and Stanton 2015), (Piwowar 2011)</td>
</tr>
<tr>
<td></td>
<td>Practices</td>
<td>(Kervin, Finholt, and Hedstrom 2012), (Vertesi and Dourish 2011), (Wicherts, Bakker, and Molenaar 2011), (Roche et al. 2014)</td>
</tr>
<tr>
<td></td>
<td>External</td>
<td>(Kervin, Finholt, and Hedstrom 2012)</td>
</tr>
<tr>
<td></td>
<td>Demographics</td>
<td>(Youngseek and Stanton 2015), (Harper 2012), (Fecher, Friesike, and Hebing 2015)</td>
</tr>
</tbody>
</table>

are not under the control of the designers. For instance, the SDR designer cannot create an intervention that would modify the age of the collaborators. Similarly, designers would not be able to create a global crisis in an attempt to modify scientists’ data-sharing behaviour. However, SDR designers should be mindful of these characteristics, design around them, and expect the relationships identified here to surface during the analysis of the collaborations. For example, designers should be expecting different behaviour from scientists at different stages of their research life cycle.

Dependent factors

Dependent factors include the perception of benefits, costs and risks associated with data-sharing. For instance, one of the major perceived costs in data-sharing is that of time-cost to prepare a dataset for reuse. This preparation includes annotating data with metadata required for the reuse of data. Recently, SDR designers have created tools that allow data annotation consistent with community standards \[\text{Wol+11a}\]. Thus, reducing the perceived and real cost of data-sharing. In subsection 4.5.2, we further explore the design features that will exploit this type of factor.
3.2.4 Conclusions

In this section, we have presented the current state of the literature with regard to factors associated with data-sharing behaviour in the Life Sciences. For the most part, the factors associated with data-sharing are independent of the SDR designer. Nevertheless, this set of factors can be used to create empirical claims to be tested in specific cases of collaborations sharing data in the Life Sciences. In section 3.3, we discuss how these theory-inspired factors can be translated to the Galaxy Groups Collaborations.

3.3 Theory Translation

In the previous section, we identified a number of factors associated with data-sharing in the Life Sciences. In this section, we discuss how these factors can be translated into a set of collaborations from the life sciences.

We examined the methodologies available to translate the factors identified and concluded that the data available was not fit to perform the translation within the collaboration without implementing a set of interventions based on these factors.

In the following, we first discuss the methodological conflicts for testing these claims in small collaborations (contribution C4 from section 1.4: The identification of a methodological limitation for the analytic approach of context-capturing). Then, we discuss the extent to which the approach to context-capturing can be used within the limitations of the methods available.

3.3.1 Methodological Limitations

Hypothesis testing of data-sharing claims in small collaborations is limited to certain methodological approaches. As we discussed in Chapter 2, small collaborations have a limited number of participants and the life-span of the team working on the collaboration is relatively short. Therefore, the selection of the method of hypothesis testing needs to consider small samples. One potential alternative to hypothesis testing in small collaborations is the case study. Flyvbjerg has argued that case studies are an effective way of testing a hypothesis, provided that good case selection is performed. Specifically, he argued that in order to perform hypothesis testing using a case
study, the identification of critical cases or extreme cases is needed. “Critical cases” can be defined as having strategic importance in relation to the general problem.

In the case of the Galaxy Groups Collaborations, translating the factors was hindered as data collection on critical cases was not possible due to the homogeneity of collaborations. All the cases in the Galaxy Groups Programme are homogenous in the parameters identified. All the cases presented a homogenous perception in terms of benefits, cost and risk. In section 3.1, we found that there is a low perception of potential benefits across the six collaborations. Similarly, perceived risk, in the form of scooping-fear and misinterpretation, was consistent among collaborations. Concerning demographics, the collected data from all the collaborations presented similar compositions in terms of gender, discipline, attitudes and age; as shown in Table 3.1. Regarding mandates, we know that they were all under the same data-sharing policy, and enforcement of data-sharing compliance was present at the same time in the six collaborations. As with regard to practices, the collaborators expressed a wish to have uniform practices across collaborations. As the analysis was performed retrospectively, we could only rely on information which was voiced by the subjects. The six collaborations we analysed had similar patterns of data production, in that they produced data collaboratively. Under these circumstances, we could not carry out a case study for hypothesis testing.

3.3.2 Opportunities

Nevertheless, the factors that are identified in the review can enable us to situate the collaborations from the Galaxy Groups Programme in the Life Science Community and narrow down the creation of interventions to those that complement the results from the exploratory study in section 3.1.

For instance, we could make the case that funding evaluation deadlines in the Galaxy Groups Programme are a similar phenomenon to community “crises” as described by Kervin et al. In section 3.1.6, we identified that the collaborations’ motivations to share data were purely extrinsic (finding Fnd13, “Sticks” only Improve Self-deposition). We reached this conclusion by examining the collaborations’ motivations to share data and observing their self-deposition patterns. We also found that self-deposition in all the collaborations were associated with funding evaluation deadlines; see Figure 3.4. Funding evaluation
deadlines are dates in the collaborations process at which funders evaluate all the metrics of the collaborations; one such metric is data-sharing in the SDR. In subsection 3.2.3, Kervin et al. [KFH12] argues that community crises, such as the influenza virus in the Genomics community, can lead to a community-wide wave of data publishing. The difference between these two cases might stem from the actors’ motivations. In the case of the Galaxy Groups Collaboration, we argue that the actors’ motivation were material, while in the Genomics community the actors’ motivations could have been ethical. However, both events (i.e., funding deadlines and crises) are similar in that they unified the behaviour of disengaged actors.

Another case is that exposing the Galaxy Groups Collaborations to perceived benefits from data-sharing could lead them to change their data-sharing behaviour. In section 3.1, we found that subjects presented a limited knowledge of other benefits that could act as extrinsic incentives (finding Fnd14, Lack of Perceived Benefits) and we made the claim that exposing collaborators to more benefits would improve data-sharing. In subsection 3.2.3, we presented the work of Kim and Stanton, who identify a perception of benefits with positive data-sharing behaviour (e.g. [KS12] and [KS15]). Therefore, this survey adds additional support to the empirical claim defined using the empirical context-capturing approach.

3.4 Chapter Conclusions

We applied the context-capturing approach to a case of small collaborations in the Life Sciences. Using the empirical context-capturing approach, we managed to identify inductively a set of factors associated with data-sharing that could be used to create empirical claims. On applying the approach, we performed an exploratory case study in which we interviewed ten subjects in six small collaborations. We complemented our observations with log data from the collaborations’ Scientific Data Repository. We found that, within small and distributed scientific collaborations from Systems Biology that are assembled as consortia the initial setup of collaboration is truly important in terms of data-sharing engagement. We found that, in small collaborations, data-sharing practices are correlated with the data and publishing life cycle and collaborators have misleading perceptions of their collaborations sharing practices. Additionally, we found that, data-sharing
behaviour is mostly motivated by extrinsic factors, partly because scientists in this collaborations do not perceive benefits in the activity. On the contrary, they perceive a great risk.

We then reviewed factors that are associated with data-sharing in the life sciences and obtained a second set of factors associated with data-sharing. In analysing the review found that most of the uncovered factors associated with data-sharing are not in the control of SDR designers (e.g. researchers’ demographics, mandates of funding policies and practices). Nevertheless, there is a sizeable set of factors that could be used to translate and create interventions (i.e., perceived benefits, costs and risks).

On translating this second set of factors to the case of the Galaxy Groups Collaborations, we encountered a number of limitations to the analytic approach. First, the factors identified in the survey were not in the control of the SDR designer, and that poses a problem for the purpose of this work. Second, the size and life span of the collaborations, combined with the historical aspect of the study, limited the methods we could use for translating factors. More importantly, translating factors retrospectively in small collaborations requires researching “critical cases”, which were not available due to the nature of the small scientific collaborations (i.e., the Galaxy Groups Collaborations). This limitation establishes an important precedent for the use of the methodology in the future for both researchers of small collaborations and SDR designers.

Nevertheless, the factors we have identified in the review in conjunction with those identified in the exploratory study can be used to create design claims for SDRs. Both types of factors represent the context of the collaboration and the discipline of our case. Therefore, the outcomes of these two context-capturing steps represent the context of data-sharing of the Galaxy Groups Collaborations.

In the next chapter, we take the set factors identified here, and we use them to define a set of interventions using the context-incorporation procedure.
Chapter 4

Persuasive Intervention Design

It’s not about what you get, it’s how you cut it, and what comes out at the other end.

Tony Zhou commenting on “F for Fake” by Orson Welles

In Chapter 3, we focused on the evaluation of the methods for capturing context to the research stage within the Scientific Data Repository development life cycle. In this chapter, we steer our attention towards the methods incorporating context to the design stage; see Figure 4.1.

In particular, we introduce the procedure to design persuasive interventions in a Scientific Data Repository. As shown in section 4.2, this procedure uses Choice-Architecture techniques to design interventions. We apply this procedure to the SDR used by the Galaxy Groups Collaborations using the set of the factors obtained from the context-capturing methodologies in the last chapter. On applying the procedure, we first synthesised the context associated factors into...
two empirical claims. Then we used the procedure to transform those empirical claims into design claims for the SDR. We implemented these design claims in an off-the-shelf SDR as design features.

To evaluate the defined features, we ran an asynchronous self-administrated web experimental study under a single scenario and two experimental conditions. In the experiment, we requested participants to perform the deposition of a fictional dataset in a fictional Scientific Data Repository. We measured participants’ data-sharing choices and compared them between the experimental and control conditions. The experimental and control conditions differed in the type of SDR design they employed. In the experimental condition, participants experienced an SDR design with Choice-Architecture techniques, while in the control condition participants experienced an SDR that fitted current designs which do not use Choice-Architecture. We consider that, as shown in subsection 4.5.5, this evaluation offers a good compromise in terms of validity and feasibility in order to test the approach suggested in this thesis to design SDRs.

We found that, as shown in subsection 4.5.3, Choice-Architecture-based interventions can affect scientists’ data-sharing behaviour in an SDR and their satisfaction on the activity of data-sharing. One of the interventions evaluated shown a statistically significant effect on the subjects’ behaviour, while the second intervention showed promising results for a future experiment. One possible implication of our results is that our procedure using Choice-Architecture could be included in the Scientific Data Repository’s design toolkit as a viable addition to the current practices to incorporate context into the design stage.

We begin by presenting the procedure to generate Choice-Architecture-based interventions. We go on to illustrate the use of the procedure by applying it to a set of empirical claims. We then present the interventions’ evaluation and the pilot study we conducted to refine the experimental evaluation and the interventions. Additionally, we present the rationale behind the selection of the asynchronous self-administrated web experimental design as the research approach for this evaluation.

Parts of the work presented in this chapter have previously been published in the following:

4.1 Background

In Chapter 2 we established the need for an approach to incorporate context to the design stage of the SDR Development Life Cycle. We argued that an approach based on persuasive design would be best-suited given the conditions of data-sharing in academia. Persuasive design is a technology design model that serves to change attitudes or behaviours of the users through persuasion and social influence, but not through coercion [Fog02]. Back then, we suggested using Choice-Architecture techniques to implement persuasive interventions in an SDR. Choice-Architecture is a set of persuasive design techniques based on the idea that changes in the decision environment can affect individual decision-making while preserving freedom of choice [TS08]. To clarify this argument, we need to unpack a number of concepts. This section introduces such concepts and contrasts a number of persuasive design methodologies with Choice-Architecture.

We have chosen Choice-Architecture, a popular and well-known paradigm on persuasive design, to support contextual change. The paradigm is a common design approach, grounded in behavioural economics and experimental psychology, and it is used to create behaviour change in subjects by modifying the context in which the subjects perform their task. Choice-Architecture has been used to design persuasive interventions in public policy [RT11] [LG88], Health [Shi10], and technology [Vis09]. Whilst Choice-Architecture has been shown to be effective in the previously mentioned contexts, there is less work on its application in the online world (e.g. [Wan+13], [HDAS08] and [Bal+11]) and in a context where the target-group are scientists (e.g. [JL12]).

4.1.1 Choice Architecture

As mentioned in Chapter 2 Choice-Architecture is a persuasive design technique devised on the idea that changes in the decision environment can affect individual decision-making and behaviour while preserving freedom of choice. The approach alters people’s behaviour in a predictable way without forbidding any options or significantly changing their economic incentives.

The Choice-Architecture design has three elements in its core: the target-group, the target-behaviour and the techniques to create the intervention [MSW11].
The target-group refers to the group of subjects to which the intervention is directed. In the case of retirement plans, (example from Chapter 2), the target-group refers to the employees in the age to select a retirement plan. The target-behaviour is generally understood as the subjects’ behaviour that the designer is aiming to encourage. For this example, that behaviour is selecting a retirement plan better suited to the retirement age. Muncher et al. \textsuperscript{[MVS15]} uses the term “decision structuring” to refer to the technique used in this case. This term encompasses the term “structuring complex choices” used by Thaler & Sunstein \textsuperscript{[TS08]}. In this case, the decision is architected to guide the subjects – the employees – to choose the defined best retirement plan (i.e., target-behaviour).

As a design paradigm, Choice-Architecture offers a range of techniques to design persuasive interventions. Muncher et al. \textsuperscript{[MVS15]} categorise these techniques in three major categories. First, Decision Information covers Choice-Architecture techniques that target the presentation of relevant information for decision making without altering the options available to the subjects. The designer can choose techniques that provide subjects with extra information to persuade to make a decision. Secondly, Decision Structure cover cases in which modifying presentation does not constitute an available option. The designer can opt for techniques that modify the arrangement of options and decision-making format. “Rearranging-options-composition” and “setting defaults” are included in this category. Finally, Decision Assistance includes cases in which the designers can provide subjects with assistance to adhere to one specific behaviour. Techniques in this category include the “provision of reminders”, “facilitation of commitment”, and “support of self- and public-commitments”.

In this work, we use these techniques to design persuasive interventions that incorporate-context to the SDR design. In the next section, we present the procedure to Choice-Architecture, as well as all its elements.

\section{Persuasive Intervention Design Procedure}

In this section, we describe the procedure we are proposing to generate persuasive interventions using Choice-Architecture. The procedure makes use of two elements: (1) the Choice-Architecture techniques and (2) the empirical claims. We describe these two elements before we go on to describe each step of the procedure.
4.2.1 Choice Architecture Techniques

In describing these techniques we use the Choice-Architecture taxonomy developed by Munscher et al. [MVS15]: which is shown in Table 4.1. However, we do not discuss all the techniques included in the taxonomy as a number of them overlap with well-established principles of HCI design. Those techniques are simplifying, using prompted choice, and increasing/reducing physical effort. Moreover, we refer the reader to Munscher et al. for a larger explanation of the techniques and the processes behind them.

Table 4.1: Survey of information systems contextualised in the Choice-Architecture framework from Munscher et al. Illustrating the processes and mechanism of Choice-Architecture in the wild.

<table>
<thead>
<tr>
<th>Category</th>
<th>Technique</th>
<th>Inf. System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Information</td>
<td>Reframing</td>
<td>Figshare, Zenodo</td>
</tr>
<tr>
<td></td>
<td>Making own behavior visible (feedback)</td>
<td>Github</td>
</tr>
<tr>
<td></td>
<td>Making external information visible</td>
<td>Researchgate</td>
</tr>
<tr>
<td></td>
<td>Providing a reference to descriptive norm</td>
<td>Academia.edu</td>
</tr>
<tr>
<td></td>
<td>Providing a reference to opinion leader</td>
<td>F1000Research</td>
</tr>
<tr>
<td>Decision Structure</td>
<td>Set no-action default</td>
<td>Figshare</td>
</tr>
<tr>
<td></td>
<td>Using prompted choice</td>
<td>N.I.</td>
</tr>
<tr>
<td></td>
<td>Changing categories</td>
<td>N.I.</td>
</tr>
<tr>
<td></td>
<td>Changing grouping of options</td>
<td>SEEK</td>
</tr>
<tr>
<td></td>
<td>Connecting decision to benefit/cost</td>
<td>Zenodo</td>
</tr>
<tr>
<td>Decision Assistance</td>
<td>Providing reminders</td>
<td>Academia.edu</td>
</tr>
<tr>
<td></td>
<td>Facilitating public commitment</td>
<td>Dryad</td>
</tr>
<tr>
<td></td>
<td>Supporting self-commitment</td>
<td>N.I.</td>
</tr>
</tbody>
</table>

N.I.: Not identified.

To illustrate the techniques, we use examples from two sources: (1) Scientific Data Repositories and (2) Content Management Systems. In both cases, we use the examples only to illustrate the interventions that could be created with the techniques, rather than as examples of the application of Choice-Architecture. As far as we know, the designers of these systems have not purposefully used Choice-Architecture. For instance, we have contacted the designers from Figshare and they have disclosed that features, that we refer to in this review, have come about via designers’ ingenuity or by request of journals. Moreover, to the best of our knowledge, the effects of these features in these specific examples have not been studied before. This survey constitutes contribution C5 from section 1.4: The
CHAPTER 4. PERSUASIVE INTERVENTION DESIGN

identification of cognitive biases in SDRs in the wild.

Making own behaviour visible

The first decision information technique is called “Making own behaviour visible”. Unlike other decision information techniques, it deals with the process of providing the user with information about her historical behaviour. To implement it, the designer should provide feedback about consequences of the user behaviour that are disconnected temporally or spatially from decision-making in the SDR. Consequences that are temporally disconnected from the decision making in the scientific process include historical records of data depositions and new depositions. For instance, in the code repository Github, as shown in Figure 4.2, code providers obtain feedback about their historical repository modifications using a dashboard on the Github landing page.

Figure 4.2: Github’s dashboard making user’s own behaviour visible with a calendar heatmap.
Making external information visible

Making external information visible is a decision information technique that provides the user with decision-making aid. This technique can be executed by providing information about the state of the repository, benefits of open data, benefits of using the repository and announcements reduction of risks and cost. For example, Figure 4.3 shows how Academia.edu, a social networking website for academics, encourages its users to deposit manuscripts in its repository. They provide information about the benefits which the user could potentially obtain by depositing manuscripts.

Figure 4.3: Academia.edu landing page after signup references a publication from PLOS ONE that associates manuscripts depositing in Academia.edu with citations boost. This way Academia.org nudges subjects into be cataloging publications on the site.
Reframing

Reframe is the last of the decision information techniques. It differs from the rest in that it modifies current information given to the target population rather than providing previously unavailable information. The information change is made to fit a particular point of view. This can be achieved through (1) Equivalence Framing or (2) Emphasis Framing.

Equivalence framing is a framing technique that consists of translating information in a logically and mathematically equivalent way. For example, the insert in Academia.edu is framed in terms of gains (Gain Frame) or terms of losses (Loss frame):

**Gain Frame:** A study recently published in *PLOS ONE* found that papers uploaded to Academia.edu receive a 69% boost in citations over five years.

**Losses Frame:** A study recently published in *PLOS ONE* found that papers that are not uploaded to Academia.edu might miss a 69% boost in citations over the next 5 years.

Emphasis Framing is another framing technique, which consists of creating a simplification of reality by focusing the target-group on a relevant aspect of an issue. For example, the data repository services Figshare and Zenodo have two different emphasis frames while sharing the same reality and the same target-behaviour. In both repositories scientists are asked and allowed to self-deposit their data – this is the “single reality”. Similarly, both services encourage users to share their data – this is the target-behaviour. Their differences stand in the framing of each service use. Each of their frames is established by the information they provide in their tag lines. Figure 4.4 illustrates the differences in the frames of both services. Figshare, in panel A, addresses the target-behaviour by focusing the users’ attention on obtaining credit as a consequence of sharing their data. In contrast, Zenodo, in panel B, uses a neutral frame.

Refer to a descriptive norm

“Refering to a descriptive norm” is a decision information technique that provides information via a social reference. Specifically, a reference derived from the target-group. These references represent social norms. When using this technique,
4.2. PERSUASIVE INTERVENTION DESIGN PROCEDURE

designers should provide references to observable behaviour of other members of the target-group performing the target-behaviour. For instance, Figure 4.5 shows how ResearchGate, a social networking site for scientists and researchers, provides the user with the social norms from your close collaborators. This takes the form of showing every time one of them registers a new publication.

Refer to an opinion leader

“Refering to an opinion leader” is another decision information technique but this one provides information via a reference to the opinion of a leader. Designers should provide the opinion on the problem of a target-group’s leader. For example, as shown in Figure 4.6, the open science publishing platform F1000Research includes testimonials from laureate researchers talking about the importance of open data and open publication.
Set no-action default

“Setting no-action-default” is a decision structure technique that focuses on changing default options in a decision setting to fit the target-behaviour better. In setting no-action-default, the designer defines default options to align them with the target-behaviour. The degree to which the default option aligns with the target-behaviour depends on the agreement of the target-group with the behaviour. First, when the target-group is homogenous, designers ought to modify default options to align them fully with the target-behaviour. For example, Figure 4.7 illustrates the how Zenodo.org, an SDR, uses no-action-default to structure its users’ decisions towards always selecting the “open access” option. On the other hand, when the target-group is heterogeneous, the designers ought to use personalised-defaults-options based on individuals’ past behaviour \cite{MVS15, Shi10}.
4.2. PERSUASIVE INTERVENTION DESIGN PROCEDURE

Figure 4.6: F1000Research landing page referring to leaders’ opinions on the subject of Open Access.

**Change categories/grouping**

“Changing categories” or “grouping categories” is a decision structure technique that deals with the composition of options. The designer ought to modify the relative attractiveness of choice options by arranging the perspective categories, allocating alternatives, partitioning options, or presenting fine-grained categories.

For example, the sharing categories from the upload form from the SEEK4science platform allows users to select up to 18 possible permission combinations (as shown in Figure 4.8). One could envision a regrouping of those options/categories to structure the decision making.

**Connect decision to benefit or cost**

“Connecting decision to benefit or cost” is another decision structure technique that aims at linking decision-making to benefits or costs that are not part of the
Figure 4.7: Zenodo.org employs a no-action default for open access in its data upload form. In this way the form nudges user to select the open access option.

traditional incentives of target-group. The designer should link data depositions to those benefits or costs over which the repository has control. For example, Figshare offers the benefit of getting a DOI for each dataset the users deposit, as shown in Figure 4.9. The DOIs have no monetary benefit but work as an incentive.

Providing Reminders

“Providing Reminders” is a decision assistance technique that seeks to create opportunities for decisions events for the target population. When using this technique, the designer should provide reminders to prompt users towards the target-behaviour at specific times. Exemplars of systems taking advantage of these mechanisms are rare in Scientific Repositories, but they are more common in other contexts. For instance, Academia.edu reminds you to visit their website every time your name has been searched in Google. Figure 4.10 shows the email
4.2. PERSUASIVE INTERVENTION DESIGN PROCEDURE

Figure 4.8: The sharing preferences setting in the upload form of the SEEK4Science system. The grouping of the current options offer great flexibility (18 combinations) but a regrouping of option could improve the users decision making.

reminder you get once your name has been searched in Google.

Support self-commitment

“Supporting self-commitment” is a decision assistance technique that focuses on providing aid to a target-group in order to adhere to commitments they made. When using this technique, the designers provide help to complete plans established by the users. For example, Dryad defines a default embargo period of one year for all data that is uploaded and provides a mechanism to make the data available at the end of that period, thus aiding the scientists in managing their commitment to open data [Dry12].
Figure 4.9: Figshare.com upload page with DOI pre-assignation on upload. DOI pre-assignation is a benefit users can get when depositing data in Figshare.

**Support public commitment**

“Supporting public commitment” is the last decision assistance technique. This technique has for its objective: Supporting the members of the target-group to make their data publicly available by when they adhere to a public commitment related to the target-behaviour. This is a technique that we have not seen implemented in information systems for science.

### 4.2.2 Empirical Claims

Empirical claims are the second element of our approach to incorporate context into the design. Empirical claims are statements that are verified with reference to objective data [MM01]. These claims have been deductively or inductively tested in the particular context of the Scientific Data Repository and thus contain essential context to the design. More importantly, empirical claims provide the core of
the design-claims the procedure generate to create the Choice-Architecture-based interventions.

4.2.3 Scenarios

In order to incorporate (1) the empirical claims about data-sharing behaviour with (2) the Choice-Architecture techniques, we make use of scenarios [CR92]. Scenarios are narratives in which one or more subjects engage in a goal-directed activity with a computer system [Car03]. In lay terms, scenarios are stories about a user using the SDR to carry out a specific task or goal. The scenarios include the activity context; for instance, the users motivations, actions and reaction during the activity.

In our approach, the context is provided by the empirical claims we identified in the research stage of the SDR’s Development Life Cycle.

4.2.4 Persuasive Design Procedure

In this section, we describe the procedure to design interventions using the techniques and empirical claims (contribution C2 from section 1.4: A procedure to create persuasive interventions in an SDR.). The procedure has three steps: (1) Reframing the SDR; (2) Selecting the Choice-Architecture Technique; (3) Identifying Scenarios. The procedure makes use of the information captured in the context-capturing stage to create the Choice-Architecture SDR, as shown in Figure 4.11.
1. Reframing the Scientific Data Repository. The first step consists of the application of a Choice-Architecture technique: Reframing. As mentioned in subsection 4.2.1, the Reframing technique aims to modify the information the current SDR provides to its users. The change in information is made to fit a specific point of view using the “Emphasis Framing” technique. As described in subsection 4.2.1, the designer ought to emphasise the aspects that are aligned with the users’ motivations. Reframing includes three substeps:

(a) Identifying the SDR’s current frame. This step can be performed by collecting and analysing the textual information present in the SDR interface.

(b) Identifying users’ data-sharing motivations.

(c) Defining a frame that would align motivations and framing.

2. Selecting Choice-Architecture Techniques. The selection of the techniques depends on (1) the empirical claims, and (2) the information available for the technique implementation. As we do not use the empirical claims in a prescriptive fashion, the designer has some space to decide its selection, but always within the limits of the techniques available and the design alternatives provided in the design-claim.

3. Generating Scenarios. Scenarios are stories about users using the SDR to
carry out a particular task or goal. In most cases, the primary activity or task in the scenarios is the uploading of datasets to the SDR. Nevertheless, the goal in this step is to produce as many plausible scenarios as possible. To address this plausible clause, three sources of information could be used to generate scenarios: (1) users’ motivations, (2) users practices and (3) the design-claims.

In the next section, we present and illustrate the procedure in action, by designing two interventions.

4.3 Persuasive Features Implementation

In this section, we describe the procedure to design Choice-Architecture-based interventions for SDRs. To perform this illustration, we selected the SEEK repository instance used by the Galaxy Groups Consortium. As previously mentioned in Chapter 3, the Galaxy Groups Consortium is an umbrella programme of a number of small collaborations are using the SEEK as their SDR. We selected SEEK because we can use the findings identified in the study of the Galaxy Groups Collaborations. These findings are pertinent to the context small collaborations in Systems Biology. Therefore, using SEEK is pertinent to the application our approach to incorporate context into a SDR. Figure 4.12 illustrates the application of the persuasive design procedure in context with the previous stage.

Figure 4.12: Contextualising the Context-Incoporation Stage. This stage transforms the findings identified in the context capturing stage into design claims for the SDR (i.e., SEEK).
4.3.1 Empirical Claims Definition

The definition of the Empirical Claims is a preliminary step in our approach. As we defined it in the previous section this definition requires reference to objective data [MM01]. The objective data is collected during the context-capturing stage in the SDR design approach. For this implementation, we use the findings identified in subsection 3.1.7 as they are particular to small scientific collaborations from Systems Biology. Moreover, all these findings are related to context.

We do not use each finding in isolation but, we synthesise them in order to create the empirical claims. For this implementation, we took all the findings related to data-sharing practices and attitudes and synthesised them into two empirical claims.

The first claim, EC1, is that, exposing researchers to the potential individual benefits of data-sharing (e.g. Open citation Boost) could potentially improve data-sharing behaviour. In section 3.1.7 we found that, scientists in small scientific collaborations did not perceive benefits from sharing their data (Fnd14). This finding was contrasted with a number of benefits identified by the community, such as Open citation Boost. Additionally, in subsection 3.2.2 we identified that, scientists’ perception of benefits is associated with changing data-sharing behaviour. Therefore, we can speculate that exposing small scientific collaboration researchers to the potential individual benefits of data-sharing could potentially improve data-sharing behaviour.

The second claim, EC2, is that, linking mandates to extrinsic incentives with risk reduction can improve data-sharing behaviour in a small scientific collaboration. In section 3.1.7 we found that, mandates (i.e., “Sticks”) are good tools to improve data self-deposition (Fnd13). We also found that, scientists, in small collaborations, prefer to share data at periods where there is less risk of getting scooped (Fnd10 and Fnd15). Additionally, in subsection 3.2.2 we identified that, when given the opportunity to restrict the access to their data for a certain period, scientists choose a short period if their data is related to articles which they have already published. Hence, we could speculate that, in a collaboration that mandates to share data upon article publication, scientists will share their data openly and promptly when they can select riskless restriction periods.
4.3.2 SDR Reframing

We performed the reframing task in three steps. First, we identified the SDR’s current frame. Then, we identified the users’ data-sharing motivations. Finally, we defined a frame that would align motivations and framing.

Identification of the SDR’s current frame

On identifying the current frame of the SDR, we collected data from all the textual sources in the SDR interface. This information included: the SDR’s User interface information (e.g. tagline, buttons and instructions) and documentation about the SDR’s goals. We then categorised all the textual data until the SDR frame emerged.

Identification of the users motivations

In order to identify the users’ motivations for sharing data, we used the motivations study of six small scientific collaborations with the Galaxy Groups consortium that we presented in Chapter 3.

SDR and user analysis results

In terms of the SDR’s frame, we found that the current design is trying to emphasise the cooperative attitude in which data is shared for the good of the community. For instance, the main tag line of the SDR is:

\[
\text{Find, share and exchange Data, Models and Processes within the} \\
\text{Galaxy Groups Consortium.}
\]

Similarly, there is a strong sense of giving data to the community. For example, the word “share” or “sharing” is constantly repeated throughout all the interface, as shown in the left column of Table 4.2.

In terms of the users’ motivations, the study from Chapter 3 suggested that the users share data under very restricted circumstances and their sharing motivations are extrinsic and, in most cases, they prefer to avoid any risks.

Discussion of misalignments

In summary, we can see that there is a misalignment between the ethical frame of the SDRs, and the scientist’s material motivation for sharing their data.
Table 4.2: Comparison of elements reframed in the experimental set-up of the SDR. The reframing was performed using Emphais Framing.

<table>
<thead>
<tr>
<th>Frames</th>
<th>Current SDR</th>
<th>Reframed SDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repository</td>
<td>SEEK</td>
<td>cite DataLabour</td>
</tr>
<tr>
<td>Name</td>
<td>Upload data: Create</td>
<td>Create Citable Data: Create</td>
</tr>
<tr>
<td>Primary Button</td>
<td>Upload and Save</td>
<td>Make</td>
</tr>
<tr>
<td>Secondary Button</td>
<td>Find, share and exchange Data,</td>
<td>Control and make your data</td>
</tr>
<tr>
<td>Tag Line</td>
<td>Models and Processes within the Galaxy Grp Consortium.</td>
<td>citable.</td>
</tr>
<tr>
<td>Instructions</td>
<td>Sharing: Here you can specify who can view the summary of, get access to the content of, and edit the Data file.</td>
<td>Select citing preferences: Who can cite/reference this data-set now?</td>
</tr>
<tr>
<td>Options Preferences</td>
<td>“Share with Public”</td>
<td>“Public can cite”</td>
</tr>
<tr>
<td></td>
<td>“Share with people involved in my projects”</td>
<td>“People involved in my project can reference”</td>
</tr>
<tr>
<td></td>
<td>“Share with my collaborators”</td>
<td>“Collaborators can reference”</td>
</tr>
<tr>
<td></td>
<td>“Nobody, keep it private”</td>
<td>“Nobody, keep it private”</td>
</tr>
</tbody>
</table>

The discrepancy between users’ motivations and the SDR’s framing suggests that the current frame may not motivate users to share data. The alternative is to define a frame that is aligned with the users’ material motivations. In Chapter 3, we identified that the perception of benefits, such as data citations, are associated with improving data-sharing. Therefore, we propose framing data-sharing in this SDR as an activity for obtaining citations. Therefore, we can claim that:

RFC: An SDR interface that frames data-sharing as an activity that potentially lead to obtaining citations

+ may promote feelings of gaining a benefit upon uploading data.

Implementation of the SDR reframing

In order to apply the frame, we exchanged all the references to data-sharing for data citing. We defined a name for the SDR that would highlight the labour
involved in producing or collecting data; this aspect was also stressed in the review of factors associated with data-sharing, as shown in section 3.2.

Figure 4.13: SEEK4Science landing page completely reframed to highlight data-sharing as a task that could lead to obtaining citations.

The reframed SDR interface using the gain frame in each aspect of the interface is illustrated in the right column of Table 4.2, we changed every piece of information previously analysed to obtain a new frame. Figure 4.13 illustrate all new interface.

4.3.3 Benefits of External Information

The first empirical claim to be analysed is claim EC1 which states that users would be more likely to change their data-sharing behaviour if they were aware of the individual benefits arising from depositing in an SDR.

**EC1**: Exposing researchers to the potential individual benefits of data-sharing could potentially improve data-sharing behaviour.
It could be said that one potential intervention for improving data-sharing would be one that provides information about the benefits of data-sharing upon data deposition to the SDR.

Taking this into consideration, we selected the best-suited techniques to the information available. We selected *Making external information visible* as the Decision Information technique to implement. With this technique, we can take empirical claims about benefits of data-sharing and use them to create the intervention. Therefore the empirical claim \( \text{EC1} \) can be re-stated as design-claim, \( \text{DC1} \), as follows:

\[
\text{DC1: An SDR interface that prompts information about the benefits of data-sharing} \\
+ \text{may promote feelings of gaining a benefit upon uploading data.}
\]

With that in mind, we used the three sources of information and brainstormed the topic of candidate scenarios under which users would share data. In the study presented in Chapter 3, we found that users visit the SDR to upload data upon (1) publishing an article; or (2) when a funding evaluation is approaching. Hence, we generated the following scenario:

**Usage scenario**: Alex shares her data in the SEEK repository. Alex, a PhD student, participating in a small collaboration, visits their data repository, upon publishing an article about her work from the last six-months. Alex is trying to upload her data into the repository. Upon getting into the upload form, Alex is prompted with information about the benefits of sharing her data. After filling the form and reading the information provided she decides to select the most accessible permissions for sharing her data.

In order to identify data that could work as “external information,” we extracted a set of individual benefits we identified in the literature review presented in Chapter 2. These include (1) Data citation from third-party dataset reuse, (2) open data citation boost, and (3) Co-authorship from third-party dataset re-user. We selected *data citation from third-party dataset reuse* as the claims stated in original publication that draw this beneficial association (i.e., Piwowar [PDF07]) gives us the opportunity to use another Choice-Architecture Technique: Equivalence Framing. In [PDF07], Piwowar states that:
Publications which share their data were cited about 70% more frequently than publications which do not. [PDF07]

We inserted the information about the benefits of data-sharing that is linked to data citation in uploading form before the users make the selection of accessibility. Figure 4.14 present the implementation of this feature.

Publications which share their data were cited about 70% more frequently than publications which do not. – Piwowar, H et al. 2007

Figure 4.14: Implementation of making external information visible. Using Piwowar claims reframed in a loss-averse fashion.
4.3.4 Embargo Period Selection

The empirical claim EC2 states that users would be more likely to change their data-sharing behaviour if they had the freedom to share their data at times that they considered to be less risky.

**EC2**: Linking mandates to extrinsic incentives with risk reduction can improve data-sharing behaviour in a small scientific collaboration.

It could be said that, over time, the data producer’s perception of data’s value and thus their perceived data-sharing risk is diminished. Few options can be considered in designing an intervention under this empirical claim. For example, Providing Reminders could be used to alert scientists to share their data a period after they have produced it. Another possibility is, supporting self-commitment by helping the data providers to decide on the less risky period in which to share their data. Along the same line, one could be establish categories/grouping to help data providers to choose sensible and non-risky periods in which to share their data. Additionally, one could use the no-action-default technique to make the selection of a future release date (data-sharing date) mandatory. For this illustration, we used the last three techniques in designing an intervention. Therefore the empirical claim EC2 can be re-stated as a design-claim (DC2) as follows:

**DC2**: An SDR interface that by defaults required the definition of a future date of data publishing

+ may promote data providers’ self-commitment to share data.
+ may reduce the perceived risk of the data provider to share their data.
+ may increase the number of open datasets in the long term.
- may put data providers off and deter them from uploading data.

We generated the following scenarios:

**Usage scenario**: Alex shares her data in the SEEK repository. Alex, a PhD student, collaborating in the Galaxy Groups Consortium, visits the SEEK repository, at which point her collaborators requested her to share the data between one another without using email. At this point, neither Alex or her collaborators has published an article
on this data. Alex is trying to upload her data into the repository. Upon getting into the upload form, Alex is prompted with information about the benefits of sharing her data. That includes information on data citation boost. After filling the form and reading the provided information she decides to give access only to her collaborators. After her selection, the repository requests her information about the future data release date. After considering the length of her project and their article’s potential publication date, she decides to select a two year period.

**Usage scenario:** Alex shares her data in the SEEK repository. Alex, a PhD student, collaborating in the Galaxy Groups Consortium, visits the SEEK repository, at which point her collaborators requested her to share the data between each other without using email. At this point, neither Alex or her collaborators has published an article on this data. Alex is trying to upload her data into the repository. Upon getting into the upload form, Alex is prompted with information about the benefits of sharing her data. That includes information on data citation boost. After filling the form and reading the information provides she decides to give access only to her collaborators. After her selection, the repository requests her information about the future data release date. Simultaneously, the SDR provides information about her collaborators’ preferences with regards to future data releases. After considering the length of her project and their article’s potential publication date, she decides to select a two year period.

We defined the grouping of categories of the embargo periods based on a number of sources. For example, six months is the expected period of data withholding for Sequencing (EMBL Nucleotide Sequence Database) and Metabolomics (MeTRO) data in the UK [BBS10]. One year is the maximum period of concealment for Arabidopsis microarray data (NASC Affymetrix service) and Crystallography (Protein Data Bank) [BBS10]. Three years after project completion is the standard data-embargo-period for a community with no defined guidelines [BBS10]. Ten years is the standard period of preservation required by scientist [BBS10]. With this information, we established four groupings for the categories: 6 months, one year, three years and ten years.

Additionally, we used the referencing to a social norm technique to create a
Figure 4.15: Design Claim DC2 represent the implementation of a combinations of techniques: establish grouping and referring to a social norm.
more persuasive intervention. We created a statement that established a norm in terms of future release dates. We used information about the collaborators contract’s length (i.e., three years) to craft this statement. Figure 4.15 shows the implementation of this intervention.

Norm: 77% of users share their data openly within three years.

In this section, we have defined two design claims (DC1 and DC2) with the Choice-Architecture based procedure for context-incorporation. We have implemented these two design claims on the SEEK4Science platform, and we use the next section to describe the evaluation we carried out to test their effects on real users. As of today, a live version of the implementation can be found online and the source code be found in Appendix B.

4.4 Pilot Study

In this section, we present the pilot study we conducted to (1) identify flaws in our experimental protocol and (2) gather rich feedback from the interventions we created. The study was of an exploratory character and consisted of a think-aloud activity and interviews with a group of researchers from the field of Biology.

We identified two potential problems during this pilot: (1) the baseline could prompt the subjects and give away the cover story, and (2) the users need more guidance to perform the experiment.

4.4.1 Design

The research questions that guided this pilot study were:

RQ1: Can the participants follow the protocol?

RQ2: What are the usability problems related to the conditions or/ and the protocol?

RQ3: What are the effects on participants’ sharing preferences?

RQ4: What are the difference between participants’ preferences (Baseline) and participants’ choices (during the experiment)?

\[ \text{http://seek.cs.man.ac.uk/} \]
**RQ5:** What’s the participants’ awareness of the difference between the conditions?

We used the think-aloud method to investigate participants’ perception of the experimental protocol and the interventions. The think-aloud method is a commonly used procedure for gathering data on usability testing in product design and development. In this method, participants are requested to say whatever comes into their mind as they complete the task.

The think-aloud sessions were conducted individually at the participants’ offices or over Skype. In some cases, we provided a laptop and allowed participants to use the browser of their preference. During the think-aloud session, we encouraged the participant to think aloud while they performed a single run of the experimental protocol. Figure 4.16 shows a description of the full experimental protocol. We manually gathered data in the form of comments and physical cues (e.g. long pauses, looking for what to click next, diversions from the protocol).

Figure 4.16: Experimental protocol as object of study of this pilot. Abbreviations: Pt.: Participant; Random.: Randomization

Our target-group was early-career scientists from (1) the field of Biology, (2) that produce datasets in electronic format. We followed a snowball approach to reach potential participants. We reached a total of 12 potential participants and enrolled 9 of them. They came from disciplines such as Systems Biology (4), Aquatic Ecosystems (2), Biotechnology (2) and Chemistry (1).

In order to evaluate the changes in the experimental protocol, we modified the experimental protocol as we were gathering data from each of the sessions. Thus each new participant experienced a slightly different experimental protocol to assess. Each iteration addressed the concerns raised by previous participants.

A list of the data and electronic analysis tools used here can be found in Appendix B.
4.4. PILOT STUDY

4.4.2 Results

In terms of the experimental protocol perception, we found that the instructions were not enough for the participants. In a number of occasions, participants forgot the instructions and sometimes they showed confusion when the protocol’s second period began. In the case of condition awareness, the first three subjects were unable to tell they were receiving different conditions in each period of the protocol. They felt that they were repeating the previous task.

If we turn now to participants’ perception of usability, we found that they were less satisfied with experimental condition A. Nevertheless, participants that first received the control condition B remained consistent on their usability perception with the experimental condition.

Although at this point, we cannot conclude anything from the intervention effects, we could observe that most participants remained consistent with their data-sharing behaviour independently of the condition they received. The few participants that did change their behaviour elaborated that their decision change was due to forgetting about the scenario. Additionally, the subjects commented on the baseline questionnaire and how that gave them an indication of the type of experiment we were conducting. Regarding the difference between participants’ preferences (Baseline) and participants’ choices (during the experiment), we found a split decision. Half of the participants were more conservative in their data-sharing behaviour in the experiment than in the baseline.

4.4.3 Discussion

Overall we found that the protocol and the instructions were not sufficient to instruct subjects in what they were doing and the reasons behind the exercise. This could have been caused by the lack of reiteration on the instructions’ presentation. To address this concern, a number of actions were implemented. First, a warning screen was added to the protocol with a reiteration of the scenario and instructions right before interacting with each condition. Additionally, an instructing overlay was added to each condition, together with a help button which provided the instructions of the activities that could be performed.

In terms of the protocol itself, the fact that the baseline survey contained questions about data-sharing preferences could have prompted participants to

\[2\text{We used the library Chardin.js https://heelhook.github.io/chardin.js/}\]
understand the nature of the experiment and thus void the experiment’s cover story. To address this, we modified the protocol and changed the order of the Preferences survey to be placed at the end of the protocol.

In the next section, we will address the execution of the full experiment using the modified protocol and a cross-over design.

4.5 Procedure and Interventions Evaluation

In this section, we present the evaluation of the interventions we designed using the Choice-Architecture-based Interventions Design procedure for SDR. The evaluation is of a functional character; we seek to test whether the procedure can produce interventions that persuade scientists (i.e., users) to change their data-sharing behaviour. To perform the evaluation, we ran a pseudo-randomised crossover-trial. In achieving this, we used the implementations of the procedure that were presented in section 4.3 as treatments.

We found that Emphasis Framing could affect scientists’ data-sharing behaviour when having to decide to share a dataset upon publication, and their satisfaction when performing the activity (contribution C3 from section 1.4). On the other hand, we ran into difficulties in recruiting participants to test all the interventions defined here.

We begin this section with the description of the experiment. We then present the results and a discussion of the implications of the procedure within the methodology to incorporate context into the SDR design, and the implications for the domain knowledge of Choice-Architecture.

4.5.1 Research hypothesis

In essence, the experiment seeks to test two hypotheses which derived from the design-claims that were used to create the interventions described in section 4.3. These design-claims are:

DC1: An SDR interface that prompts information about the benefits of data-sharing

+ may promote feelings of gaining a benefit upon uploading data.

DC2: An SDR interface that by default required the definition of setting an established period of data embargo
4.5. PROCEDURE AND INTERVENTIONS EVALUATION

+ may promote data providers’ self-commitment to share data.
+ may reduce the perceived risk of the data provider to share their data.
+ may increase the number of open datasets in the long term.
- may put data providers off and deter them from uploading data.

RFC: An SDR interface that frames data-sharing as an activity that potentially lead to obtaining citations
+ may promote feelings of gaining a benefit upon uploading data.

The hypotheses we test in this experiment are:

H1: An SDR’s interface designed to exploit their users’ interests on obtaining citations will not affect their data-sharing behaviour (in terms of data-sharing accessibility choices)

H2: An SDR’s interface designed to reduce their user’s perception of risk will not affect their data-sharing behaviour (in terms of embargo period’s length)

In the next section, we present the experimental design we used to test these two hypotheses.

4.5.2 Experiment

Here we present each element of the experimental study which we ran to test the interventions created with our procedure. We begin with a description of the experiment. Then we present the experimental design, subjects recruitment strategies and statistical analysis plan. Finally, we present a description of metrics we used.

In brief, we ran an asynchronous self-administrated web experimental study under a single scenario and two experimental conditions. In the experiment, we requested participants to perform the deposition of a fictional dataset in a fictional Scientific Data Repository. In the scenario, we told participants that they were part of a European scientific collaboration. They had just published an article, and they have been requested to deposit their primary data in a Scientific Data Repository. In essence, they follow the scenario we presented in section 4.3.
Additionally, we disguised the experiment as a web usability experiment to avoid potential psychological biases. We asked participants to evaluate the usability of two SDR while we collected their data-sharing preferences.

**Scenario:** Alex shares her data in the SEEK repository. Alex, a PhD student, collaborating in Galaxy Groups Consortium, visits the SEEK repository, upon publishing an article about her work from the last six months. Alex is trying to upload her data into the repository. Upon getting into the upload form, Alex is prompted with information about the benefits of sharing her data. That includes information on data citation boost. After filling the form and reading the information, she decides to select the most open accessibility permission for sharing her data.

We measured participants’ data-sharing choices and compared them between the experimental and control conditions. The experimental and control conditions differed in the type of SDR design they employed. In the experimental condition (A), participants experienced an SDR design with Choice-Architecture techniques while in the experimental condition (B) participants experienced an SDR that fitted current designs that do not use Choice-Architecture.

In summary, the evaluation presented here aims to (1) test the Choice-Architecture-based interventions against current SDR implementations and (2) evaluate the functionality of our procedure. The latter is achieved by comparing the Choice-Architecture based interventions against a set of features that in theory should give completely different results.

**Design**

We used a 2x2 crossover study design for the experiment. A crossover study design is a well understood experimental design in which a sequence of measurements is gathered from each participant under changing treatment conditions [KJ94] [WB12] [Mil+09]. We selected this setup for two reasons. First, we had clear and accessible metrics with which to establish a quantitative study (i.e., the length of embargo periods and accessibility choices). Second, with an experimental setup, we can perform a rigorous assessment of the Choice-Architecture techniques. Previous studies have shown that additional sources of information (such
as experiencing a website) can diminish the effects of the Choice-Architecture interventions [LSG98] [HDAS08].

In our study, each participant performed a remote session on a single day with the control and experimental condition in counterbalanced order. They performed it on a computer of their choice over a web-browser.

The selection of a self-administrated web study (asynchronous) and the crossover design (contra-balanced order) was informed by concerns over the participants’ availability driven by our target population and methods of recruitment. Our target population was early-career scientists in the field of biology which (1) limits the participants pool and (2) we consider they would be less likely to participate in a study that would require both spatial and temporal availability due to their work restrictions. Our method of recruitment – web invitations – has been previously associated to have low response [SF08] [Nul08]. The participation rates we considered (e.g. [Nul08] [MAV05] [SF08]) indicated that the response rate would be in the range of 20–40%. We confirmed this concern as we encountered a 20% response rate to our invitations. These two limitations drove the selection of our study selection.

The asynchronous experiment has both advantages and disadvantages that we addressed. An asynchronous experiment has the benefit of (1) reducing the spatial and temporal barriers [And+07] and (2) reducing the artificiality setting – that being their computer at their offices. On the other hand, the asynchronous study poses three problems. First, we were not able to capture richer information about participants’ interaction with the interventions. Thus, we used a synchronous pilot study to capture user experiences during the experiment and made adjustments in the setup – see section 4.4. Second, we had no control over the participants’ demographics within our target population – i.e. their field of science, English proficiency, age. Therefore, we used pseudo-random assignment through hashing [BE13a] to randomise the participants based on the demographic information they gave in the initial survey. Finally, participants on asynchronous studies can be prompted to give careless responses. Thus we addressed careless responses with a self-reported screening item and bogus screening item [MB12] in the exit survey – as shown in Appendix C.

The crossover design has a number of advantages the most important for this study was that it needs half as many participants compare to a parallel experimental design [Che+97] [Tra10]. This design leads to more powerful tests than
simply comparing two independent groups using between-participant information. As each participant acts as his control, between-participant variation is eliminated as a source of error \[\text{TS09}\]. Furthermore, this design has been previously used in a number of other experiments using the same type of interventions \[\text{McG+99}\]. However, the use of this design has been criticised in a similar context due to the use of improper length washout periods, which, diminishes the results of crossover study \[\text{McG+99}\]. We addressed that concern by providing both a washout period and including an interruption activity that addresses the effects of the interventions – more details about the washout period can be found in section 4.5.2.

![Figure 4.17: Full Crossover Trial Protocol.](image)

We conducted the experimental protocol shown in Figure 4.17. First, the participants were given the scenario. Next, a six-items demographics survey was administered to the participants – refer to Appendix C. Then, each participant received the control (B) and experimental (A) condition in counterbalanced order. The experiment provided the participants with a fictional dataset to use within either condition. After each condition, the participants were asked to fill a usability evaluation (ASQ) for the condition they experienced. We administered a memory interruption activity during the washout phase to eliminate the carry-over effects. Finally, an eight-item exit survey, of their data-sharing preferences, was administered to the participants before debriefing.

The modified version of the SEEK4Science platform used for this experiment \[\text{Owe+16}\] can be found in Appendix B.
4.5. **PROCEDURE AND INTERVENTIONS EVALUATION**

**Scenarios and Experimental Conditions** We used a single scenario for this experiment. The scenario consisted of four aspects: the type of collaboration the participants belong to; the step in the data life cycle participants are in the scenario; extrinsic incentives towards and against data-sharing. In the scenario, participants were requested to pretend that they belong to a European scientific consortium. In the scenario, participants were told they were at the publication release stage in their data life cycle and their incentives to deposit their data were funding body mandates.

The experimental condition (A) aimed to modify two target-behaviours: the data access selection and the data embargo period selection. In data accessibility selection, we attempted to modify the participants behaviour towards selecting more open preferences on the data they deposited. In the data embargo period selection, we tried to modify the participants behaviour towards selecting a shorter embargo for the deposited data. To modify those target-behaviours, the condition (A) was created using the procedure and included a number of Choice-Architecture techniques. The full detail of the application of the procedure can be found in section 4.3. The Choice-Architecture techniques used in this implementation included: (1) reframing, (2) changing categories, (3) making external information available, and (4) reference to a norm.

The experimental condition (A) and control condition (B) are similar and contrasting in a number of aspects. In condition (B) both decision events and target-behaviours were prompted in the same way that we did with condition (A). However, the differences between conditions stand in the Choice-Architecture techniques we used. Condition (B) does not include the optimal frame in the application of the reframing technique. The frame for condition (B) was selected to conform to current policies in Biological Sciences and SDR implementations. Moreover, condition (B) lacks the implementation of any other Choice-Architecture techniques. A detailed description of the control (B) can be found in Appendix C. The differences between experimental condition (A) and control condition (B) strive from three reasons. First, by not using the optimal frame in the application reframing, we can compare the Choice Architecture-based Intervention procedure functionally at different levels. Second, by not using the Choice-Architecture techniques, we can test both the interventions and the procedure against other standard procedures and standard interventions that use no Choice-Architecture techniques.
Both conditions were implemented on an instance of SEEK4Science platform[^3].

**Cover Story**  We used a request for a user usability evaluation as the cover story in the experiment. The cover story was administered in the initial instructions and revealed in the debrief. We used the after After-Scenario Questionnaire (ASQ) [Lew90] as the mock instrument the participants were using to evaluate the website. Additionally, we included an extra item to measure the satisfaction or regret the participant had over their decision. Despite the fact that these items were used only with the purpose of the cover story, we collected data from this evaluation to compare the participants perceived usability of the experimental condition (A) and the control condition (B), as well as their regret rankings. The usability of the treatments was estimated with a cumulative link mixed model and assessed by a likelihood ratio test. The regret observations were assessed using Wilcoxon rank sum test.

**Washout Activity**  After performing the task for the first period of the 2x2 crossover design, the participants were presented with an activity that aimed to induce a memory interruption. This is the washout. In the interruption activity, the participants were involved in the classification of rotated and mirrored characters. It has been showed that this type of exercise impairs the participants’ attention shifting and short-term memory encoding [Pan+11]. We applied this activity using the set of characters defined by Young [YPL80] for a similar experiment.

**Subjects Recruitment**

Our target population was early-career scientists from (1) the field of Biology, (2) that produce datasets in electronic format and (3) have worked under an open data-sharing policy in a certain point of their career. In order to attain the target population, potential participants were invited through speculative invitations (n=10), using top-down requests from Principal Investigators (n=14), advertising in research volunteering networks and social media. We used three instruments for the invitations: two templates for invitations that can be found in Appendix C and a poster that can be found Appendix C[^3].

[^3]: Seek4Science is an SDR Framework for the field of Systems Biology http://www.seek4science.org/
Figure 4.18: Participants breakdown. Showing a 20% response rate.
The recruitment campaign had a reach of 268 potential participants and had a 20% response rate. A total of 54 participants were enrolled. Participants were excluded from the study if they did not belong to the field of Biology, did not complete the experiment, gave random answers in the survey questionnaire or lacked English proficiency [Cos+14] [KHA12]. A total of 40 participants were used for the analysis. Figure 4.18 shows details of their participation.

**Carryover Effects**

We carried out a carryover effect assessment and confirmed that the experimental design was not skewing the observations. The carryover effect is a characteristic error effect of the crossover study design; it occurs when a treatment effect carries over to the next period [WB12]. The carryover effects were assessed following the two-step Grizzle approach with a Wilcoxon rank sum test on differences of the medians between periods for three different measures [WB12] [Che+97].

In the results the accessibility measures, we observed an indication of a potential positive carryover effect – see Figure 4.19. However, the statistical analysis did not identify a significant carryover effect between the sequences when measuring data accessibility ($p = 0.3009$). Moreover, we carried out an analysis in terms of usability, see Figure 4.20 and we did not identify a significant carryover effect between the sequences ($p = 0.9422$). Therefore, we are confident that the design is not biasing accessibility and usability observations.

![Figure 4.19: The effect on the accessibility means over the two periods seem to be positive. Showing a non statistical significant carryover effect.](image)

In terms of the regret observations, we observed an indication of sequence group effect – see Figure 4.21. This is a borderline significant carryover effect.
between the sequences ($p = 0.0688$). We confirmed this sequence group effect after performing a crossover analysis and controlling for participants choices (OR=0.43, CI=[0.20,0.93]). Therefore, we only used data from the first period of the study to analyse regret observations.

Figure 4.20: Effect on the usability means over the two periods. Showing a non statistical significant carryover effect.

Figure 4.21: Sequence group effect on the regret means over the two periods. Showing a border line significant carryover effect.

**Statistical Analysis**

The effect of the experimental condition on the participants’ preferences was estimated with a cumulative link mixed model and assessed by a likelihood ratio test. The direction of the estimated treatment effect was estimated with and Odds ratio estimation. We consider the Cumulative Link Mixed Model (CLMM)
as the fittest method of analysis as it has been suggested to fit our design (fixed and random effects) and the type of data we collected (Ordinal data). CLMM has been used previously used to analyse a similar data \cite{Ran89} by Christensen R. \cite{Chr15} and by Kuyumcu et al. \cite{Kuy+14}.

Differences between populations experimental sequences were assessed by Wilcoxon rank sum test for numerical variables and Pearson’s Chi-squared test for categorical variables.

A list to the data \cite{Gar16b} and electronic analysis tools \cite{Gar16c} used here can be found in Appendix B.

**Metrics**

To evaluate the changes in data-sharing choices, we measured both dataset accessibility and embargo period length. The data accessibility ranking consisted of a fourth level ranking: private, collaborators, stakeholders, everybody. The groupings were inductively created from the case study presented in Chapter 2. The embargo period ranking consisted of five-level ranking: 0 months, six months, one year, three years, ten years. subsection 4.3.4 contains more information about the metrics design. To measure usability for either condition we used the after scenario questionnaire by Lewis et al. \cite{Lew90}.

**Sample size estimation**

We estimated a sample of 46 participants. Assuming a crossover trial with \( n \) subjects will have about the same power as a parallel trial with \( 2n \) subjects \cite{Jul04} \cite{Pia05}, we estimated the sample for a parallel trial with an estimated power of 90\%, \( \alpha = 5\% \) and \( OR = 3.0 \) \cite{Tra10}.

**4.5.3 Results**

In general, our results have shown that we have collected data from the target population, as presented in section 4.5.2, that is, (1) early-career scientists (2) from the field of Biology, (3) that produce datasets in electronic form and (4) have worked under an open data-sharing policy in a certain point of their career. Our sample included scientists from a number of branches of Biology; more than half of the participants were data producers from the field of Bioinformatics (60\%).
Table 4.3: Experimental Sequences Comparison and Participants demographics. Showing a balance between sequences due to the pseudo-randomisation.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sequence A-B</th>
<th>Sequence B-A</th>
<th>A-B ≠ B-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (median)</td>
<td>38 years</td>
<td>40 years</td>
<td>p&gt;0.05</td>
</tr>
<tr>
<td>Labour (median)</td>
<td>31 hrs</td>
<td>42 hrs</td>
<td>p&gt;0.05</td>
</tr>
<tr>
<td>Discipline (top 3 only)</td>
<td></td>
<td></td>
<td>p&gt;0.05</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Systems Biology</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Biochemistry</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Collaboration size (n)</td>
<td></td>
<td></td>
<td>p&gt;0.05</td>
</tr>
<tr>
<td>1 member</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2-4 members</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5-12 members</td>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>&gt;20 members</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Sharing Preferences</td>
<td></td>
<td></td>
<td>p&gt;0.05</td>
</tr>
<tr>
<td>After Publication</td>
<td>17</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Before Publication</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Update Data</td>
<td>8</td>
<td>8</td>
<td>p&gt;0.05</td>
</tr>
<tr>
<td>Store Data Openly</td>
<td></td>
<td></td>
<td>p&gt;0.05</td>
</tr>
<tr>
<td>every 6 months</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>every 12 months</td>
<td>2</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>never</td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.4: Comparison of Accessibility Results. Using three different CLM Models. Showing the effect of the Emphasis Frame intervention.

<table>
<thead>
<tr>
<th></th>
<th>No Treatment</th>
<th>Treatment</th>
<th>Treatment wControls</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence</strong></td>
<td>1.98</td>
<td>1.64</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>[0.30; 12.86]</td>
<td>[0.19; 14.15]</td>
<td>[0.05; 16.15]</td>
</tr>
<tr>
<td><strong>Period</strong></td>
<td>2.23</td>
<td>1.96</td>
<td>2.35</td>
</tr>
<tr>
<td></td>
<td>[0.85; 5.88]</td>
<td>[0.70; 5.54]</td>
<td>[0.59; 9.36]</td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td></td>
<td>3.28∗</td>
<td>5.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[1.01; 10.62]</td>
<td>[0.91; 28.73]</td>
</tr>
<tr>
<td><strong>AIC</strong></td>
<td>98.71</td>
<td>95.59</td>
<td>92.90</td>
</tr>
<tr>
<td><strong>BIC</strong></td>
<td>110.62</td>
<td>109.88</td>
<td>111.55</td>
</tr>
<tr>
<td><strong>Log Likelihood</strong></td>
<td>-44.35</td>
<td>-41.79</td>
<td>-38.45</td>
</tr>
<tr>
<td><strong>Num. obs.</strong></td>
<td>80</td>
<td>80</td>
<td>76</td>
</tr>
<tr>
<td><strong>LR.stat</strong></td>
<td>5.1187∗</td>
<td>6.1373*</td>
<td></td>
</tr>
</tbody>
</table>

OR ∗ 1 outside the confidence interval.  
Confidence intervals in brackets.  
LR.stat ∗∗∗p < 0.001, ∗∗p < 0.01, ∗p < 0.05

Most of them were in their late 30s and all of them produced data in electronic form.

Additionally, we can say that the sample represents participants from (1) small size collaborations, (2) that prefer to share data after publishing results, (3) have data that requires less than 50 hrs of work to be produced and that (4) do not follow the best practices on data depositing. More than half of the participants collaborate with a small scientific team, and only a few do their scientific work on their own. On questioning them about their data-sharing preferences, the large majority said that the best moment to share a dataset is at publication milestone. Twenty percent of the participants said that they don’t use data repositories, and more than half of them said that they never update datasets once they uploaded them to a repository.

In terms of randomisation of our sample, we found non-significant differences (p > 0.05) between the groups in the two experimental sequences when it comes to the field of science the participants belong to, age, collaboration size, their data-sharing preferences, labour involved in the data they produced, and their use of data repositories. Full details can be seen in Table 4.3.
4.5. PROCEDURE AND INTERVENTIONS EVALUATION

Figure 4.22: Data accessibility preference comparison under two conditions. Control condition (top) and Treatment condition (bottom). The participants under the Treatment condition selected more open data accessibility options for the data they uploaded. The response is given in terms of the accessibility setting selected.

**Accessibility**

In terms of accessibility, we observed that the treatment improved the accessibility performance in comparison to the control. In the descriptive statistics, we observed a slight increment in accessibility preferences on participants under the treatment – as shown in Figure 4.22. We corroborated that change when we found a significant treatment effect on the accessibility choices (LR Stat = 5.1187; p=0.02367). The results, as shown in Table 4.4, indicated that using the treatment increases the odds (OR= 3.27) of selecting a more open accessibility option. Furthermore, the treatment effect was still present when controlling for other variables. We found that these effects remain to a lesser degree when controlling for participants discipline and labour invested in the datasets (LR Stat = 6.1373; p=0.01324). Complete results are shown in the third row from Table 4.4.

**Embargo Periods Length**

In terms of embargo periods selection, the treatment created slight improvements, but we did not find statistically significant effects. We observed, as can be seen in
Figure 4.23: Aggregated participants embargo Periods preference comparison under two conditions. Control condition (top) and treatment condition (bottom). Participants under the Treatment condition (A) selected shorter embargo periods for the data they uploaded in the experiment. Response is given in terms of the length of the embargo period selected.

Figure 4.23 a reduction of embargo period’s length by the treatment. However, we could not assess its significance as not enough participants were exposed to this intervention.

Usability evaluation

For the usability evaluation, we found, as shown in Table 4.5, that the control received better usability rankings. There was a significant treatment effect on the usability ranking (LR Stat = 4.4; p=0.03481). The results indicate that participants give lower usability rankings to the treatment condition (OR= 0.22).

Regret

Although the participant showed more regret on their choices under the control, we found no relationship between the regret and the participants future choices. There was a significant treatment effect on the regret ranking (W = 292, p= 0.009651, CI=[0.00004, 1.99]), but we did not find a participants choices effect on the regret measurements (p=0.6178).
Table 4.5: Comparison of the Usability Results. Using the Cumulative Link Mixed Model.

<table>
<thead>
<tr>
<th></th>
<th>without Treatment</th>
<th>with Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>[0.03; 22.89]</td>
<td>[0.02; 39.21]</td>
</tr>
<tr>
<td>Period</td>
<td>2.51</td>
<td>2.87</td>
</tr>
<tr>
<td></td>
<td>[0.67; 9.45]</td>
<td>[0.62; 13.19]</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.04; 1.09]</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>157.71</td>
<td>155.25</td>
</tr>
<tr>
<td>BIC</td>
<td>174.38</td>
<td>174.31</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-71.85</td>
<td>-69.63</td>
</tr>
<tr>
<td>Num. obs.</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>LR.stat</td>
<td>4.4544*</td>
<td></td>
</tr>
</tbody>
</table>

OR * 1 outside the confidence interval.
Confidence intervals in brackets.
LR.stat ***p < 0.001, **p < 0.01, *p < 0.05

4.5.4 Discussion

Accessibility

We found that a Scientific Data Repository with an interface framed to emphasise the self-interest aspects of data-sharing (i.e., obtaining citations from data re-users) could positively affect scientists’ data-sharing behaviour in terms of accessibility. Emphasising self-interest could increase the sense of long-term expectations of the users of SDRs and thus make them select higher accessibility permissions for their data.

Our findings support previous findings from the Choice-Architecture paradigm in the online context. For example, our findings are in agreement with Mako Hill’s findings on his study of online Encyclopaedias [Mak10] in which he argues that the success of Wikipedia could have been influenced by the frame of the Online encyclopaedia. Similarly, these results further support the idea that the inclusion of framing statements in web user interfaces influences user’s judgment as shown by Hartmann et al. [Har+08] in their study of the presentation of information on websites.

However, the fact that the treatment effect was moderated when controlling for additional variables raises new questions. The moderation when controlling for participants’ labour suggests that these factors are might be more relevant for
data-sharing behaviour. Users providing datasets that have required a considerable amount of time and labour into their creation would become less affected due to the large investment put on those assets, as Borgman has suggested [Bor11]. Therefore, the potential gain offered by the frame might not be comparable with the value given to the dataset. The reduction of the effect when controlling for discipline could be caused by the over-reliance of some disciplines on primary data. Not all disciplines rely on the use of secondary data. Therefore, primary data-only users would see little gain on contributing to an SDR, because using secondary data is not a common practice for their discipline.

Embargo Periods Length

In the case of embargo periods’ length, the results are encouraging, but they need to be judged with a larger sample. One reason we believe most of the participants do not get to experience this intervention is because the scenario does not present an ideal opportunity for embargo period selection. Besides the results shown here, we conducted the same experiment under other scenarios with similar results in terms of obtaining participants to run the experiment. Therefore, new scenarios should be attempted to statistically test this intervention.

Regret and Usability

We found that participants regret their accessibility choices more in the control condition. Essentially, participants expressed less regret when the goal of the activity was the pursuit of individual gain than when they did for the community or ethical reasons. The goal of the activity, which is determined by the frame, seems to be causing the regret ranking differences. Interestingly, their accessibility selection did not affect their regret ranking. It remains unclear why the choices in itself did not modify their regret rankings.

We found that the participants ranked the control condition with higher usability ranking. In other words, we found the goal of the activity was to obtain an individual gain the usability rankings were lower. It is unclear whether there is a relationship between the participants’ decisions and the usability of the interfaces. Surprisingly this was the same result we obtained in our pilot study; see section 4.4. This effect could be due to the amount of information the participant received during the experiment. For instance, the treatment condition included more textual information than the control condition. This extra information was
4.5. PROCEDURE AND INTERVENTIONS EVALUATION

provided in the form of the application of two Choice-Architecture techniques: *Making external information visible* and *referring to a social norm*. A previous study by Ivory and Hearst [IH02] highlighted a correlation between the amount of text on a website and usability rankings. We consider this difference minimal, but the user might have a different perspective.

Strengths & Limitations

There are some noteworthy strengths and limitations of our approach. On the one hand, this is the first attempt to test the use of persuasion on changing scientists’ data-sharing behaviour. Our approach, using a web experiment, is less susceptible to biases such as subjects trying to adapt to the norm and framed questionnaires. Finding the hypothesised effect is encouraging, as framing effects are often diminished when an additional stimulus is provided in this type of setup [HDAS08]. On the other hand, our findings should be interpreted with care. For example, even when we tried to provide a realistic scenario (using an off-the-shelf widely-used and pre-populated SDR), the artificiality of the setting may have produced unnatural behaviour that might not be observed under similar conditions. Additionally, although we did not find a significant carryover effect, it is difficult to ignore the effects of memory on the participants’ responses. Furthermore, the generalisation of the outcomes should be made with care as most of the participants were from the field of Bioinformatics and as it has been shown that controlling participants by discipline does lessen the effect of the treatment. Therefore, we should pursue future lines of research. Carrying out this experiment as an A-B test in a production environment would be useful to overcome the artificiality of the setting and remove the potential effects of our crossover design. In terms of confirming the results on discipline-based differences in data-sharing behaviour, we ought to study other examples of SDR with a diversity of data providers such as in our study. The small size and short lifespan of the small scientific collaborations context would indicate that, with a number of studies of such nature, we can then later provide results with a meta-analysis.
Conclusions

We found that Choice-Architecture-based interventions could affect scientists’ data-sharing behaviour in an SDR and their satisfaction on the activity of data-sharing. We found that we could create effective interventions by incorporating context using our procedure to create Choice Architecture-based interventions. One possible implication of our results is that our procedure using Choice-Architecture could be included in the Scientific Data Repository’s design toolkit as a viable addition to the current practices. Independently of the use of our procedure, an implication of these findings is that SDR designers would need to pay special attention to the type of data and discipline of the users before designing interventions.

Nevertheless, although the procedure uses well-established techniques from software development (such as scenarios and claims analysis) we are unable to answer questions on the subject of designers’ adoption. Therefore, a study of the usability of the procedure should be pursued.

4.5.5 Methodological Rationale

Besides the experimental crossover study, a number of other approaches were considered and attempted to answer the research questions posed in this chapter. Those alternative studies include a survey and experiment on the wild. This section describes these studies and their limitations in order to answer the research questions examined in this chapter and the problems we encountered while piloting them. Figure 4.24 shows an overview of this discussion.

First, we contemplated the idea of conducting a survey to evaluate the response of the users to potential interfaces. However, a survey would not be examining the participants’ behaviour but their attitude. Therefore, we discarded this approach.

Second, we considered performing an experiment on the wild in the real repository. Bakshy et al. has shown how to perform this type of experiments and has provided tools to help in their design. Moreover, Kraut et al. have shown that this is a common approach to evaluate interventions in web interfaces. However, the selected SDR had policies in place that deter the administrators and designers to perform experiments in the running instance. Moreover, an experiment in the wild could take many months to perform. The
4.5. PROCEDURE AND INTERVENTIONS EVALUATION

Figure 4.24: Decision tree on the alternative studies considered to the experimental crossover study. The study conducted in subsection 4.5.2 is highlighted in green. The studies piloted are highlighted in yellow.
dependent variables are related to participants research lifecycle which could be in the range of 6-9 months.

A better compromise was to perform a synthetic experiment using a parallel design. A synthetic experiment is a scientific method that has been used to evaluate empirical studies in the software engineering research and practices [Ibr+11]. The evaluation design has been gaining popularity in computer technology evaluations in the last ten years [Zel09]. However, in a parallel design, we faced the problem of low participation. Following the experience of a previous study, we estimated a response rate of only 20%. The later would put a burden on obtaining enough data. Therefore, we opted for a crossover design, which allows us to halve the study’s sample size estimation [Jul04] [Pia05].

Finally, another decision we made in the experimental design was the arrangement of the interventions. As mentioned in subsection 4.5.4, the interventions were nested within the experimental design. This offered the opportunity of using the same subjects to evaluate all interventions. However, it also meant that the responses in the first intervention would deter some participants from the experiment with the second intervention. This led to a large number of participants only experiencing one intervention. Therefore, we ran the experiment under two additional scenarios. These scenarios gave different starting conditions to the participants that we predicted would make participants experience both interventions. Nevertheless, none of the scenarios were able to steer enough participants to the second intervention (i.e., embargo period).

4.6 Summary

In this chapter, we have presented a this thesis’s procedure to incorporate context to the design stage of the Scientific Data Repository development life cycle by using Choice-Architecture-based interventions. We grounded our work in the Choice-Architecture paradigm. We have described the steps of our procedure: Reframing the SDR, Selecting Choice-Architecture Techniques, and Generating Scenarios. We have shown the procedure application to two empirical claims. To evaluate our procedure, we have run an experiment and shown that we can successfully create interventions that by incorporating context modify data-sharing behaviour.

In the next and final chapter, we draw together conclusions for our work in
incorporating context to the SDR design and propose an agenda for future work.
Chapter 5

Conclusions

This study aimed to know whether the design of the SDR can influence science data sharing behaviour from small collaborations. The general literature on this subject, and the literature in the context of small collaborations, is inconclusive on several questions within the scholarly communications discourse. This PhD research sought to answer some of these questions.

Making research data publicly available is a movement that is building momentum. In many ways, it is the best opportunity to gather the range of observations that are necessary to accelerate discoveries. However, this paradigm poses a conundrum for scientists in small collaborations, who are carving out their niche and have limited resources. Scientists, in this context, disagree about how much, and when, they should share data. Hence following best practices, such as sharing data in a scientific data repository, are not commonly followed in small scientific collaborations.

Scientific Data Repositories (SDR) have emerged as a platform that can help small collaborations to facilitate science data-sharing. Today’s environment in academic research has led to underutilisation and under adoption of SDRs, presenting designers and researchers with the challenge of identifying design features that effectively contribute to the effort of open science data.

The work presented in this thesis proposes an approach that supports designing and studying interventions that would facilitate data-sharing. These contributions aim to support Scientific Data Repository designers and small scientific collaboration researchers. We based the core of this work on the notion of including contextual knowledge into the SDR design methodologies as an effective strategy in facilitating data self-deposition and data-sharing in science.
The inclusion of context into the design methodologies allows designers to consider variables that are not captured through the requirement gathering stages. Taking into account that science data self-deposition and data-sharing in SDRs is not widely adopted ([Cra+10], [Ten+11a], [Ten+15], [Esc11], [Roc+14], [JB14]), we developed an approach for capturing and incorporating the context of data-sharing into the design of Science Data Repositories.

In this concluding chapter, we summarise the research contributions that have resulted from this investigation. Figure 5.1 supports this discussion and illustrates the structure of the thesis we have presented.

Figure 5.1: Structure of the thesis presented

5.1 Thesis Summary

In the course of this thesis, we presented and evaluated an approach to facilitate data-sharing. The approach facilitates data-sharing by including the data-sharing context into the Scientific Data Repository design. First, we investigated the scientific community approach to data-sharing. We conducted this investigation
at two levels. At one level, we examined the literature on practices, attitudes and factors associated with scientific data-sharing; see Chapter 2. At second level, we examined the literature on factors related to scientific data-sharing within the field of Life Sciences: see Chapter 3. We conceptualised data-sharing as a decision-making problem for which designers need to include more context in the SDR design: see Chapter 2.

We reviewed the current approaches to SDR design and the SDR development life cycle. To achieve this, in Chapter 2, we conducted a review and synthesis of the SDR design practices. We identified the steps required to tackle the problem in the development life cycle and the current design approaches’ limitations to context-capturing and context-incorporation.

Once the groundwork was laid down, we defined the approach that facilitates data-sharing in a self-depositing SDR. To achieve this objective we combined two approaches. First, we proposed a combination of empirical and analytic socio-technical approach to capture context. Second, we proposed using Choice Architecture techniques for incorporating context into the design. We presented an overall description of the procedure in Chapter 2 and the specifics in Chapter 3 and Chapter 4.

Finally, in Chapter 3 and Chapter 4, we evaluated the effectiveness of the mentioned approach. We applied the approach to a group of collaborations in the Life Sciences using an SDR and ran an experimental evaluation in an SDR setup. In applying the approach we interviewed the collaborations’ participants, we studied their SDR, their norms and their behaviour. In Chapter 4, we demonstrated that we could incorporate that context into design claims and design interventions. In the experimental evaluation, we have shown that interventions created using this approach can effectively impact data-sharing behaviour.

5.2 Empirical Findings and Implications

The main empirical findings are summarised within the respective chapter: in Chapter 2, Chapter 3, and Chapter 4. Our findings are not only of interest to researchers studying data sharing; they also inform choice architecture researchers and SDR designers by providing insights into opportunities and challenges of new approaches to design scholarly communication systems. This section synthesises the empirical findings to answer the Thesis’s research questions.
How can Science Data Repository design influence data-sharing behaviour?

We found that the design of the SDR could influence users’ data sharing behaviour. Our results indicate that the SDR design can modify the users’ perception of individual benefits and risks at the moment of depositing data in the SDR. In subsection 4.5.3 we have experimentally shown how an SDR designed using persuasive design techniques can predictably improve data sharing behaviour. Behaviour change using web interventions has been shown to work in other contexts, as we show in section 2.6. The observed correlation between the SDR design and positive data sharing behaviour might be explained in this way: when the SDR design and the users’ individual motivations are aligned, users would more likely behave as the design is intended. Although persuasive design is used in many different contexts, there are no documented cases of its use to influence scientists’ decision making. As shown in section 2.6 most documented design approaches do not make use of persuasion as a tool for SDR design. On the other hand, persuasive design is sometimes employed in a non-prescripted fashion as shown in our review of the information system for science in subsection 4.2.1.

The principal theoretical implication of this research, for SDR designers, is that the modification of the SDR’s user interface can have serious implications for the behaviour of scientists using the SDR to preserve, collaborate or publish their data. Additionally, the evidence from this research suggests that the persuasive design model can be extended to a population previously unexplored in the Choice Architecture literature (i.e., Biologist) under a previously explored decision-making problem (i.e., Social Dilemma). Another important practical implication is that these finding suggest that research data management teams could also invest in the SDR design to address data withholding in their collaborations. At the moment, research data management teams around the world rely mostly on “sticks” to address research data withholding (i.e., policies and mandates). Unlike, policies and mandates, the SDR design modifications as describe in this research, would require less negotiation with collaborators than the implementation of a new policy.

Which type of these theories are the best suited candidates to be translated?

We found that Economics theories concerning individual benefits, costs and
risks are a good descriptive model to explain scientists’ data sharing attitudes and practices. In section 2.5 we discussed the current misalignment of cost, risks and benefits in the open data movement. Furthermore, in section 3.2 we distinguish a number of studies that identify associated factors with data sharing and categorise them in terms of risks, costs and benefits. In subsection 3.1.7 we found that this argument holds true for small collaborations. It seems possible that these results are related to the notion of data being a currency in science, and that currency is incredibly valuable particularly in small collaborations. The results of the literature survey that we conducted in section 3.2 seem to be consistent with the argument we presented in this work.

These findings have three important implications for scholarly communication researchers. First, researchers in this area can make use of Economic models in order to study data withholding problems. Moreover, scholarly communications researchers can expect that explanatory theories of data-sharing in other fields of science could be, in some specific scenarios, generalised from one collaboration to the other, even when the type of collaboration is different in size. Taken together, these findings suggest that scholarly communication researchers need to identify new ideas to address the perceptions of risk and individual benefits that scientists hold.

What type of techniques can we effectively use to transform empirical claims into design interventions?

We found that Choice Architecture techniques could be effectively used to design thriving features for SDRs. In section 2.7 we show that the data sharing problem fits the Choice Architecture model. Additionally, in subsection 4.2.1 we have shown that some SDR implementations are already using some of the cognitive mechanisms that underpin Choice Architecture. In section 4.3 we have shown how these techniques can be easily implemented in a data repository. Two facts may account for the effectivity of these techniques in this particular context. First, the techniques are grounded in well-founded psychological processes and mechanisms, as shown in section 2.7.2. Second, in our approach we combined them with standard design practices from software design, as shown in subsection 4.2.4. As previously mentioned, although no comparable uses of Choice Architecture techniques have been found in the literature of scholarly communications, we have found SDR implementations in the wild that take advantage
of the processes and mechanisms underpinning these techniques, as shown in subsection 4.2.1.

These findings can be used by Choice Architecture researchers that seek to identify cases of the use of cognitive biases to influence scientists’ decision-making in the web. Moreover, SDR designers can use the Choice Architecture models to evaluate the designs and features they include in their SDR in a predictive fashion. A further implication for scholarly communication researchers is that they could use the Choice Architecture supporting theories as a framework to explore the data withholding problem.

How can practitioners incorporate the techniques we provide into their design process?

We found that the proposed procedure could be used to incorporate the persuasive paradigm effectively into the design of SDRs. The procedure used Choice Architecture techniques and tools borrowed from software development. In subsection 4.2.4 we presented the design approach and its deployment. Moreover in section 2.6, we have shown how our proposal differs from others’ approaches to design SDRs.

The most obvious implication for SDR designers is that our results suggest that this procedure can enhance their toolkit. We can also foresee some practical implication for researchers of small collaborations. For example, we recommend that scholars of studying small collaborations should prioritise the identification of critical cases in their research as early as possible to avoid the pitfalls of only having successful and visible small collaborations in their studies.

5.3 Contributions

In this section, we summarise the main contributions of the results of this work.

C1: A set of factors associated with data-sharing behaviour. In section 3.2 we conducted a review of the state of the art data-sharing associations. We restricted this review to the Life Sciences. The resulting set of factors from this survey contribution can be used by practitioners (e.g., SDR designers or Open Science Data researchers in the Life Sciences) to understand that variables are at work in data-sharing. In the review, we point out that the majority of data-sharing
associations in the literature are out of the control of the SDR designers and that a very limited number of interventions have been proven to improve data-sharing behaviour.

C2: A procedure to create persuasive interventions in an SDR. In section 4.2, we presented and evaluated the procedure to create design claims for an SDR using Choice-Architecture techniques. We have shown that, using the procedure, we are able to create a set of design claims that can be built into SDR features and we have shown that the features effectively improve data-sharing behaviour in an experimental environment. Practitioners can use the results of this methodological contribution to: (1) create interventions in Scientific Data Repositories; and (2) define features in SDR development frameworks.

C3: We found that Emphasis Framing can help to modify data-sharing behaviour. In section 4.5, we evaluated a design intervention based on the use of Emphasis Framing and found that it can effectively improve data-sharing behaviour. This empirical contribution can be used by Choice-Architecture researchers as another example of the using of Emphasis Framing in decision problems classified as social dilemmas with expert subjects.

C4: Identified a methodological limitation for the analytic approach of context-capturing. In section 3.3, we apply this thesis’s proposed approach to context-capturing to a set of collaborations from the Life Sciences. We carried out an exploratory study on the collaboration and a survey of data-sharing associations. We aimed to apply our analytic approach to context-capturing but identified a number of limitations in its retrospective application to small scientific collaborations. The results of this methodological contribution should warn practitioners wishing to design SDRs for small scientific collaborations on two aspects. First, on using an analytic approach for context-capturing, practitioners need to proactively identify “critical cases” for hypothesis testing. Second, on identifying factors associated with data-sharing in the literature, practitioners should aim to identify factors that are actionable by the SDR, instead of factors that are out of the control of the SDR design. Under other circumstances, we do not
recommend using the analytic approach as the sole context-capturing tool.

**C5**: Identification of the use of cognitive biases in SDRs. In section 4.2 we conducted a survey of information systems for scientists that rely on the use of processes or mechanisms of Choice Architecture. This survey can be used by practitioners to identify naturally occurring experiments in the area of Choice Architecture with expert subjects.

### 5.4 Future Work

There are several directions that we hope to take this work in the future.

**The context incorporation procedure evaluation.**

In Chapter 4 this thesis’s evaluation of the procedure to create interventions focused on the functional aspect of the interventions. Although we based the procedure in common design practices [CR92], it can be expected that designers/practitioners with no previous knowledge of Choice Architecture would find the techniques complicated to use. Therefore, our evaluation would benefit from a usability evaluation of the application of the procedure.

**Evaluation of further persuasive interventions needs to be explored.**

In this thesis, we proposed a Choice-Architecture based intervention for data embargo selection. Our preliminary evaluation of the intervention appears to be promising, but the full analysis is needed in order to for it be completed. As of January 2016, the Data Repository service Figshare began to offer a similar design feature in their service [Hah15]. We have talked with the designers and confirmed that they came up with the design through different means than those presented here. More importantly, Figshare has a pool of users larger than any of the experiments we could devise. Therefore, we are currently exploring the necessary steps with their design team to evaluate the feature in a production environment via a natural experiment.
Strategies to improve participation in SDRs.

We recommend studying design features that would produce sufficient added value to change scientist’s adoption behaviour. In this thesis, we proposed an approach to facilitate data-sharing in SDRs but our approach is limited to modifying the behaviour of active-users. However, a challenging unaddressed problem in scientific data-sharing is the poor levels of adoption of SDRs’. There have been attempts to draw scientists to adopt specific SDRs by adding features that increase the value of data deposited. For instance, designers have implemented features such as credit mechanisms \[\text{Ope15}\], utilities such as metadata cataloging \[\text{Wol+11b}\] and data annotation features \[\text{Wol+11a}\]. Is important that scholarly communications researchers study the effectivity of these features how their adaptation into an SDR can improve data-sharing without causing “motivational misalignment”.

Exploring the application of the full range of Choice Architecture techniques

We suggest studying and evaluating the full range of Choice Architecture techniques in different cases. So far, we have explored only two techniques to address the problem of data withholding. We speculate that techniques such as “providing reminders” and “supporting self-commitment mechanisms” can be crucial to improve SDR adoption and change data sharing behaviour. As we have shown in subsection 4.2.1 these techniques have been used in the domain to engage users with other scholarly communications systems. Researchers and designers will need to identify the use cases in which these techniques can be applied and generate the right interventions.

Strategies to improve the study of Small Collaborations.

We suggest that further work is required to identify the methodologies to study small collaborations. In this work, we proposed the use of an analytical case study, but we encountered a number of methodological drawbacks, as shown in subsection 3.3.1 When studying, these small groups researchers would need to be creative on identifying small collaborations. We speculate that looking at registries of public-funded grants is a viable strategy to identify them (e.g. Gateway to Research in the UK, http://gtr.rcuk.ac.uk/). Additionally, we speculate that
researchers would need to be proactive in engaging critical cases to test hypotheses. As these cases are essential for the evaluation of theories in these studies.

5.5 Final Thoughts

In this work, we show that we can facilitate data-sharing by incorporating context into the design of a Science Data Repository and identified a set of restrictions to use in our approach. The approach proposed in this thesis can be used by practitioners wishing to improve data sharing in an SDR, while contributions such as the survey of factors associated with data-sharing behaviour can be used by researchers to understand the problems related to data sharing in small scientific collaborations.

I hope that this effort, together with those by funding bodies, journals and the open science community, will allow all of us, as a community, to reach a stage in which data-sharing in small science is a reality. Moreover, I hope this research thesis will contribute to a deeper understanding of data-sharing behaviour and the more efficient design of Scientific Data Repositories.
Bibliography


[LSG98] Irwin P. Levin, Sandra L. Schneider, and Gary J. Gaeth. “All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects”. In: *Organizational Behavior and Human Decision Processes* 76.2 (Nov. 1998), pp. 149–188. ISSN: 0749-5978. DOI: 10.1006/


[PDF07] Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma. “Sharing Detailed Research Data Is Associated with Increased Citation Rate”. In: PLoS ONE 2.3 (Mar. 21, 2007), e308. DOI: 10.1371/journal.pone.0000308. URL: http://dx.plos.org/10.1371/journal.pone.0000308 (visited on 10/16/2012).


[Ren+12a] Yuqing Ren et al. “Building Member Attachment in Online Communities: Applying Theories of Group Identity and Interpersonal Bonds”. In: MIS Q. 36.3 (Sept. 2012). 00038, pp. 841–864. ISSN:
0276-7783. URL: http://dl.acm.org/citation.cfm?id=2481655.2481665 (visited on 03/28/2014).


[RIN08] RIN. To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. 2008. URL: http://eprints.soton.ac.uk/266742/


[TP11] Anne E. Thessen and David J. Patterson. “Data issues in the life sciences”. In: ZooKeys 150 (Nov. 28, 2011), pp. 15–51. ISSN: 1313-2989. DOI: 10.3897/zookeys.150.1766. PMID: 22207805. URL:


[Vis09] Arun Vishwanath. “From Belief-Importance to Intention: The Impact of Framing on Technology Adoption”. In: Communication Monographs 76.2 (2009), pp. 177–206. ISSN: 0363-7751. DOI: 10.1080/


Appendix A

Interview Guide

1. How challenging was to adapt to the work styles of partners from other institutions?

2. Could you compare the chemistry you have your collaboration at your institution with that of the collaboration in the project.

3. When making decisions, how often than not was your opinion considered seriously?

4. Could you recall an occasion in which the partners got together to solve a problem related to the mission of the collaboration? How were they solved?

5. Can you explain the roles and responsibilities of you and your institution in the collaboration?

6. Have you been present in meetings related to the management and about how the collaboration is run? How would you rate those meetings in terms on making the collaboration work well?

7. Can you talk about the specific goal of this collaboration/project? Have there been discrepancies with other members about this goal?

8. Could you talk about how the collaboration/project organise to work smoothly?

9. What’s your mission as a researcher? In which degree participating in the collaboration hinders you from meeting your mission?

10. How much is your independence affected by having to work with the collaboration?

11. What is the collaboration expecting from you? What your institution is expecting from you? To what degree is here a conflict between these expectations?

12. To what extent did you combine resources with the other partners?
13. Your organisation shares information with partner organisations that will strengthen their operations and programs?

14. Has there been any type of appreciation from the other members for what you brought to the collaboration?

15. Would you have reach the results that you got if you were not in the collaboration?

16. Could you recall an instance in which disagreement between partners raised? How it was solved?

17. What types of assets would you trust to other members of the collaboration?

18. Overall, why did you consider was best to remain in the collaboration rather than leave it?

19. Can you talk about how were you and the partners communicating while the project was running?

20. Can you talk about your experience using other repositories to get data.
Appendix B

Supplementary Materials

Interview Transcripts and Survey Responses

All my outcomes are related to my ORCID: http://orcid.org/0000-0003-3484-6875.


Data from Pilot study


Data from online experiment


Source code for data analysis of transcripts and analytics


Source code for statistical analysis of the experiment


Experimental instrument implementation

Appendix C

Experiment Supplementary

C.1 Demographics Questionnaire

1. What is your primary subject discipline?

2. What type of materials (research data) do you produce/collection?

3. In what year were you born?

4. In terms of your English speaking skills, which of the following conditions applies to you:
   - I’m a native English speaker
   - I have lived in an English speaking country for more than 2 years
   - I have been living in an English speaking country for at least 10 months
   - I’ve learnt English when I was a child (> 12 years old)
   - I have an English proficiency above professional working level
   - You can select more than one.

5. How many people are involved in the process of creating your materials?
   - only me
   - between 2 and 4
   - between 5 and 12
   - between 13 and 20
   - more than 20
   - (that could include for example other researchers, students, or lab technicians)

6. Can you estimate how many days of work per person are needed to produce/collection one of yours materials?
C.2 Exit survey

1. I have never attended an academic course in my entire life.

2. Which website did you trust the most to perform the activity you performed?
   - First one
   - Second one
   - Some sentences appeared in the upload form in the websites you visited, can you identify any of them?
   - 77% of the community makes their materials citable within 3 years.
   - 69% the community shares their materials
   - Publicly available data is significantly associated with a 69% increase in citations.
   - None of the above

3. In which way(s) do you share digital material with researchers outside of your institution?
   - when I am requested for the data
   - by uploading it in an open repository
   - by uploading it in my (or my institution’s) website
   - I have never shared materials openly

4. When is the best time to make your materials available to the scientific community?
   - +10 years after publication of a peer-review paper
   - +5 years after publication of a peer-review paper
   - +3 years after publication of a peer-review paper
   - +1 years after publication of a peer-review paper
   - immediately after publication of a peer-review paper
   - before publication of a peer-review paper

5. How frequently do you upload materials for open access?
   - every month
   - every three months
   - every six months
6. Do you ever change or update materials you have previously uploaded?

7. In your honest opinion, should we use your responses in our analyses in this study?

C.3 Motivations Questionnaire

1. To which Project do you belong?*Required

2. What is your scientific background?*Required What’s your role in your Project
   - Modeller
   - Experimentalist
   - Bioinformatician
   - Other

3. If you DO upload assets to SEEK, why do you do it? What motivated you to upload these assets? NOTE: Assets could be a data-set, a model, a SOP.

4. If you DO upload assets to SEEK, Who has access to the assets you uploaded? Who has permission to view and download your assets?
   - The members of my project
   - The whole Galaxy Groups Consortium
   - They are publicly available to the world
   - Only me
   - Other

5. If you DO upload assets to SEEK. Why did you choose the level of access you decided to select?

6. If you DO upload assets to the SEEK. Which sentence better defines your reasons for uploading assets to the SEEK
   - I contributed with assets to the SEEK due to contractual obligations
   - . . . because I like the idea of the SEEK
   - . . . because I think it’s useful to make these assets available to other people
• . . . in order to obtain further funding for my project
• I contributed with assets to the SEEK in order to preserve my assets for the future.
• . . . because I want to increase the visibility of my research
• . . . in order to facilitate the research process with my collaborators

C.4 Email Invitation

Dear * * * * *

I am writing to you in your capacity as member of the * * * * * Consortium and I am doing so upon encouragement by * * * * * from the University of Manchester where I am currently completing PhD in Computer Science.

I would like to request your support as we would like to get your insight about your work in the * * * * * project and your experience with * * * * * SEEK. The purpose of the interview is to identify the ways members organised themselves to collaborate towards building a collection of assets in * * * * * SEEK, how did they deal with conflict, and the alignment between their goals and objectives.

The interview would be 45 minutes long and it will be recorded. Interviews would be scheduled at your convenience (phone, Skype or in-person). All your identifying information would be treated as confidential, it will not be revealed in the aggregated findings.

Should you be willing to participate, please accept this invitation here or contact Dr. Katy Wolstencroft (katherine.wolstencroft@manchester.ac.uk) , or email me at kjg@cs.man.ac.uk.

Thank you very much for your consideration and we are looking forward to hearing from you. Your participation will contribute to the improvement of collaboration features of SEEK and to my PhD research at University of Manchester.

CLICK TO ACCEPT INVITATION

Yours sincerely,

C.5 Usability and regret items

1. Overall, I am satisfied with the ease of completing the tasks in this scenario.

2. Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.

3. Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks.

4. I regret the access preferences I’ve chosen in the website
C.6 Poster Invitation

Figure C.1: Poster and flyer invitation used to recruit participants.

C.7 Control Condition
Figure C.2: Control condition’s landing page. All the textual information was
framed to fit the current frame of the SDR.
C.7. CONTROL CONDITION

Figure C.3: Control condition’s upload form. Two intervention were added in this design, an Emphasis Frame for the accessibility options and we avoided to use grouping and classify for the embargo period selection.