Understanding and Improving Human Data Relations

Alex Bowyer

Contents

Additional 1: Data Protection Terminology and a Legal Definition of Personal Data **2**

Additional 2: Sentence Ranking - Sentences and Analysis **3**

Additional 3: Storyboarding Action Cards **4**

Additional 4: Notation for Quotations in Chapter 4 **7**

1 Additional 5: Family Civic Data Categories **8**

Additional 5: The Private Data Viewing Monitor **9**

Additional 6: GDPR Data Analysis Approach **9**

Additional 7: Best and Worst Companies for GDPR Handling **18**

Additional 8: BBC R&D's Cornmarket Project **19**

Additional 9: Hestia.ai, and Sitra's *digipower* Project **21**

Additional 10: DERC's Healthy Eating Web Augmentation Project **22**

Additional 11: Special Attribution Note for Chapter 7 **23**

Bibliography **24**

Additional Reference Information

Additional 1: Data Protection Terminology and a Legal Definition of Personal Data

From the GDPR (Hoofnagle, Sloot and Borgesius, 2019) and its antecedents, a number of concepts have been established which are relevant to this thesis, specifically (Information Commissioner's Office, 2014; The European Parliament and the Council of the European Union, 2016):

- *Personal data* is legally defined as any information relating to an identifiable natural person - one who can be identified directly or indirectly by reference to an identifier such as a name, identification number or location or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.
- The *data subject* is the identified individual, living or deceased, who the personal data relates to.
- A *data controller* is the legal entity (company, public authority, agency, individual or other body) which collects or stores personal data about an individual and determines the means and purposes for which it is processed. Liability for data protection compliance rests with the data controller.
- A *data processor* is a legal entity (company, public authority, agency, individual or other body) which deals with personal data as instructed by a controller for specific purposes and services offered to the controller that involve personal data processing.
- *Personal data processing* refers to any manual or automated handling of digital or analogue data including collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.
- A *Subject access request* is the right to a copy of your personal data.
- *Data portability* is the right to receive a copy of all stored data about you, not just that which you provided, in an accessible and machine-readable format such as a CSV file, so that you can transport it to another service or make use of it.

The terms *Subject Access Request* and *Data portability* are used in Case Study Two, and referenced also in Chapter 7.

For simplicity, this thesis uses everyday layperson-friendly terms rather than the legal terms defined in this section. Data subjects are referred to simply as **individuals** and both data controllers and data processors as **data holders**, because for this thesis, focusing as it does on the individual perspective, there is no need to draw a distinction between data controllers and data processors.

Additional 2: Sentence Ranking - Sentences and Analysis

In this section, additional details are provided on the *Sentence Ranking* exercise referenced in 4.2.6.

The sentences offered to participants across the 4 workshops were as follows:

- `S1` A family's data should all be joined up and looked at together.
- `S2` Any information from more than 5 years ago should be hidden from staff.
- `S3` Asking families for consent to share data just once at the start is enough.
- `S4` Councils should treat families like people, not records in a database.
- `S5` Families don't want to be responsible for looking after their data.
- `S6` Families find setting privacy preferences to be annoying and tedious.
- `S7` Families should always be able to talk to someone from the authorities about their data.
- `S8` Families should have rights to see their data and how it is used.
- `S9` Families will be willing to spend time checking their data is correct.
- `S10` Families won't mind lots of data being collected about them if they can see it.
- `S11` Families' data should be private unless they say it can be shared.
- `S12` Information stored about families must be fair and accurate.
- `S13` It is important for support workers to know mental health details.
- `S14` Just looking at data doesn't tell you everything about a family.
- `S15` Labels like 'domestic abuse' are damaging to families & hard to shake off.
- `S16` Numerical scores are a good way to compare the progress families have made.
- `S17` Officials should be able to see historical records about families.
- `S18` Public sector officials can make good judgements just by looking at families' data.
- `S19` Support workers make better decisions if they have more data about a family.
- `S20` Support workers should be able to see family medical records.
- `S21` The police should be able to see all of a family's data.

Where participants unanimously or mainly disagreed with a sentence, it is referenced in the inverse using a prime notation, e.g. `S18'`, which would imply a reference to the opposite of the statement - in this case 'Public sector officials can **not** make good judgements just by looking at families' data.'

In each of the workshops, families ranked the sentences according to:

- (a) whether they agreed, disagreed or were neutral on that statement, and
- (b) whether or not they felt that statement was important.

This produced numerical ranking data which was analysed as follows:

1. Sentence rankings were encoded on two scales. Sentences which contained a negative statement were inverted so that disagreement with them could be considered as agreement with a positive statement.
    a. *Agreement*: neutral (0) -> agree (+1.0)
    b. *Importance*: not important (0.0) -> important (+1.0)
2. Rankings from different groups within workshops were aggregated, using mean averaging, with a weighting to ensure each workshop contributes equally regardless of attendance.
3. This gave four values for each sentence, for each participant group (families only, staff only, and combined). *Variance* can be understood as 'unanimity of opinion': i.e. variance 0.0 indicates total agreement and 1.0 would indicate disagreement.
    a. *Mean agreement*
    b. *Variance of agreement*
    c. *Mean importance*
    d. *Variance of importance.*
4. Prioritising variance in agreement over variance of importance, the four dimensions were reduced to three to allow a visualisation to be produced.

The resulting visualisation is shown in Figure 4.1.

Additional 3: Storyboarding Action Cards

Drawing from the world of film production, storyboarding is a well-established technique in participatory design (Spinuzzi, 2005; Moraveji *et al.*, 2007). Usually it involves the participants drawing out a series of sketches in the form of a comic strip 'telling the story' of an interaction, encounter or activity. However, it had already been determined, both in terms of the research approach of this thesis [3.2.2], and in terms of responding to participants [4.2.6] that it would be more important to understand the interpersonal interactions between family and support worker and the actual actions performed upon or with data, rather than the mechanisms by which the data interaction would occur. Focusing on the visual aspects of information visualisation could be distracting. Therefore, I developed a novel technique for use in the phase 2 workshop: **Storyboarding Action Cards**. Each storyboard card denotes a possible action that can be carried out by a family member (yellow border), support worker (blue border) or an action performed together (green border). Each card includes a simple action summary such as 'Give Information' and an iconographic representation of the action, along with a short description of which actor is doing what. It includes blank lines which the participant can 'fill in' to describe the specifics of this occurrence of the action.

Figure 1: Figure ARI.1: Extract of Sample Scenario Storyboarding Exercise walkthrough

Based on the accumulated knowledge of Early Help processes amongst myself and SILVER colleagues, enhanced for this purpose through consultation with a former social worker, I developed a total of 43 different cards to represent the suite of possible actions that would be interesting to track. These are grouped into eight different types of card:

- **Conversation Cards** – representing actions relating to exchanges of information in a conversation as well as discussions, decisions, and questions;
- **Consent Cards** – representing actions relating to acquiring, revoking or changing family consent to data sharing or storage;
- **Data Access Cards** – representing the searching, browsing, reading, requesting and storing of information;
- **Motivation Cards** – for representing the internal wishes of either family member or support worker;
- **Feeling Cards** – for representing the emotional state of either family member or support member (This included a blank emoji face which could be filled in as well as describing the emotion in words);
- **'Elsewhere' Cards** – for those actions performed by either actor outside of their support engagement, such as sharing information with or obtaining information from a third party;
- **Problem Cards** – to represent actions where either party experiences a problem, for example either party having an issue with information handling or content, or a disagreement between worker and family member; and
- **'Custom' Cards** – a catch-all for any remaining actions that do not fall into one of the above categories.

The intent behind the storyboarding action cards is that they serve as both a boundary object and *things to think with* (as with the Family Data Cards

described in (Bowyer *et al.*, 2018)) to provoke discussion among participants. They have an additional function over the Family Data Cards, however: they can be arranged in a sequence, much like a storyboard or comic strip, and filled in, to tell the story of exactly who would do what and how in the process of a support conversation involving shared data interaction. In this way they lend themselves to model processes rather than object design. Figure ARI.1 shows an example of three cards having been filled in and arranged in sequence to tell a simple story of how a scenario of a worker seeking out an address following new information from the family member.

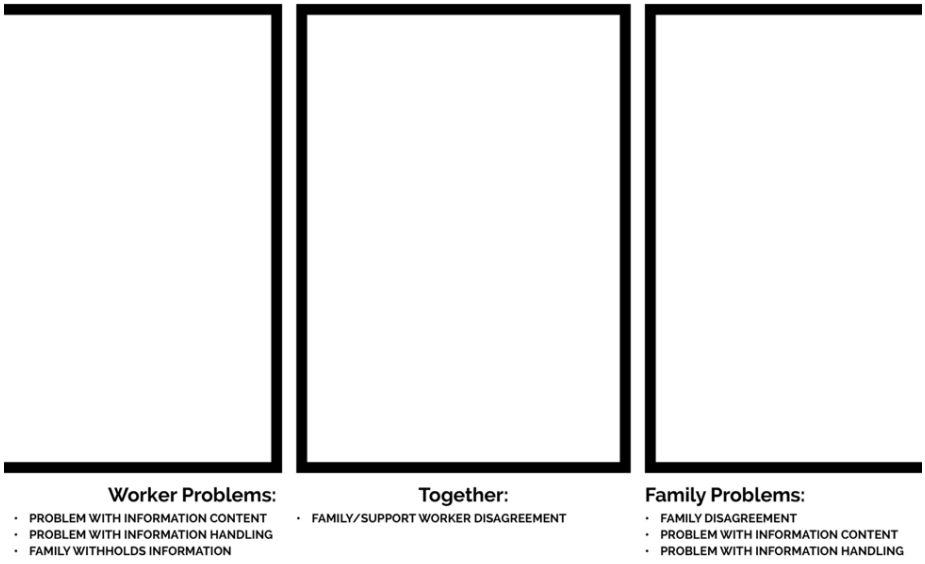

Figure 2: Figure ARI.2: Example Backing Mat for Storyboard Decks (pictured here: backing mat for all three 'problem' card decks

In addition to the storyboard cards, I also designed 'backing mats' for each of the eight card types. These were printed on large coloured card corresponding to each card type's backing colour, and provided areas for the 'decks' of available cards to be picked from. Each backing mat provided a separate home for family member actions, staff actions, and joint actions. Additionally, each backing mat included a summary of the available action cards of this type, and a prompt question. An example of a backing mat, in this case for Problem Cards, is shown in Figure ARI.2.

Introduction and Practice

In order to familiarise the participants with the storyboarding action cards and

the available actions, participants were first presented with an introduction to the storyboarding concept, as used in film-making and participatory design, then the card design and intended usage was explained.A very simple scenario of a family going through a breakup was used to talk through an illustrated example of how to map out the subsequent worker/parent conversation using the action cards. Then participants were invited to use the same scenario and practice mapping out the scenario themselves; however, this time they were to map out a 'problematic' version of the scenario, where things do not go so smoothly.

Scenario-Based Storyboarding Discussions

After the participants were acquainted with the cards and had practiced the storyboarding method, the main activity began, to which was allocated the majority of the time in the session. This involved each group mapping out two stories for a more substantial scenario; one version where things go smoothly and another 'negative' version where things do not go smoothly. It was highlighted to participants that the aim was to identify what would or should happen at each stage, and why.

The scenarios used for this activity by the two groups were (a) a new scenario where a couple is looking at their historical medical records (which contain various matters of concern such as missed appointments and historical mental health issues) and (b) a 'labels and judgement' scenario that had been used in the phase 1 workshops. Additional scenarios were prepared but not used. The layouts of the completed storyboards were photographed for reference, and to provide context during analysis of the discussion transcripts.

For a completed storyboard layout example, see Figure 3.10.

Additional 4: Notation for Quotations in Chapter 4

Quotations included in section 4.3 are references using the following notation:

- FQnn = Family Quote - a quote from the families-only workshop (A)
- SQnn = Staff Quote - a quote from a staff-only workshop (B)
- CQnn = Combined Quote - a quote from the combined workshop (C).
- Sn = Sentence n - a sentence from the *Sentence Ranking* exercise, see Additional 2.

The number after FQ/CQ/SQ provides a unique identifier for each quote, which can be used to look up the referenced quote in [TODO INSERT REF TO APPENDIX SECTION HERE]. Individual speakers are identified only by their role. Within each quote, or in brackets afterwards, the speakers are identified as *Worker*, *Parent*, *Child*, or *Researcher*.

1 Additional 5: Family Civic Data Categories

The table below illustrates the types of *family civic data* identified in the pilot study [3.4.1; Bowyer *et al.* (2018); Appendix A], and referenced in Case Study One [4.2.1].

Table 1: Table ARI.1 - Example Categories of Family Civic Data.

Category	Type of data	Examples/Details
Family	Personal details	Date of birth, address, telephone number.
	Relationships	Marital status, exs, step-parents, living arrangements.
	Children	Parentage, adoption, fostering, childcare.
Education	School Records	Attendance (truancy), special needs.
	Academic Results	SATs, reports, exam failures, training courses.
Welfare	Social Support	Social worker visits & notes, details of family crises, interventions, allegations.
	Welfare Benefits	Jobseeker's Allowance, child support, Disability Living Allowance, tax credits
Money/Work	Family Finances	Salary, savings, credit cards, spending, debt
	Employment	Job history, periods of unemployment, performance at work, NI, PAYE, pensions.
Civil	Housing data	Council house provision, eligibility criteria.
	Legal documents	Birth / marriage / death certificates, citizenship /immigration status, work permits.
Crime	Criminal records	Arrests, cautions, offenders' registers, prison time, speeding tickets, spent convictions.
	Court orders	Restraining orders, lawsuits, custody, ASBOs.
	Domestic Violence	Allegations made, medical records, social / legal interventions, victim support.
Medical	GP records	GP's notes, prescriptions, tests, referrals.
	Hospital records	Operations, hospital stays, emergency care.
	Medical conditions	Diagnoses, diseases, allergies, blood type.
	Mental health	PTSD, breakdowns, depression, sectioning.
	Addictions	Substance abuse, gambling, rehab, crime.

Category	Type of data	Examples/Details
Leisure[1]	Library Usage	Books/CDs borrowed, computer access.
	Sports & Health	Gym usage, class attendance.
	Shopping Habits	Loyalty cards, store & online purchases.
	Transport Data	Buses used, ANPR tracking, walking patterns.

Additional 5: The Private Data Viewing Monitor

By removing the filter layer on an old monitor and modifying cinema IMAX glasses, a monitor was created that only allowed viewing by the holder of the viewing glasses, which would be ideal for interviewing someone about their data while respecting privacy. Face to face interviewing had to be abandoned due to COVID-19, so this technique was sadly never used in practice.

Additional 6: GDPR Data Analysis Approach

In this section, the methodology used for the analysis of data from Case Study Two is explained. The content of this appendix is identical to Appendix 3 in the Supplemental Materials of the CHI 2022 paper from this study (Bowyer, Holt, *et al.*, 2022). Case Study Two was written first as a paper and then expanded to produce Chapter 5. While the paper was co-written, Chapter 5 was written entirely by Alex Bowyer.

All coding was carried out by Alex Bowyer and Jack Holt, who followed the following process over a nine-month period, comprising at least 200 person-hours:

1. **EXTRACTION AND ANALYSIS OF SEMI-QUANTITATIVE DATA**: Identifying closed question (or brief) responses that might be processable quantitatively.
2. **TEXT FILE PROCESSING**: Splitting, organising, anonymising and some cleaning of auto-transcribed and time-coded text files.
3. **CATEGORISATION INTO CSVs**: Categorised extraction of time-coded text sections from text files into cells of 6-topic spreadsheet, then generation of CSV files for importing into Quirkos Cloud (Daniel Turner, 2014)

[1]Some leisure categories (namely Shopping and Transport) were included that are not strictly civic data, as these would be useful for exploring issues around ethics. These also provided a reference point for participants to better consider the 'big data' benefits of data linking.

Figure 3: Figure ARI.X: Private Data Viewing Monitor with Viewing Glasses

4. **INDUCTIVE CODING**: Importing of CSVs into Quirkos Cloud and labelling by Participant, Company, and Topic. Inductive coding of source texts, ensuring good coverage per topic and per participant.
5. **REDUCTIVE CYCLES**: Reductive cycles of merging, renaming and reorganising the codes hierarchy, resulting in 10 top-level codes with hierarchies of coded texts underneath them.
6. **THEME IDENTIFICATION & QUOTE EXTRACTION**: Construction of 3 paper-focussed themes using Workflowy (Turitzin and Patel, 2010) and quote gathering using the organised codes hierarchy.

Some additional detail on the stages:

1. Semi-Quantitative Data Extraction & Analysis

Prior to beginning coding the data, responses to some key closed questions from the transcripts were combined with field notes, response emails from companies forwarded by participants, sketches and tables from Interview 1/2, data from the interview 2/3 spreadsheet cells, and other data collected, and used to populate a spreadsheet that featured summaries of those responses. For example, where participants had been asked to outline their hopes for the outcomes of their GDPR data requests, these responses were recorded on the spreadsheet to be used as a resource for summarising participant hopes in a manner that could be easily quantified and referred back to. In some cases this data was analysed within the spreadsheet to produce insights, graphs and percentages. Such data was later used to support and illustrate findings from the coding process. This spreadsheet also included important information relating to each participant's GDPR process experience, such as the timeliness and completeness of their data returns, which could serve as a reference point when analysing the transcripts.

The semi-quantitative data areas captured or derived from captured data were:

- Company Response Timelines
- Power Scores
- Trust Scores
- Hopes, Goals and Imagined Uses
- Term Definitions
- For each participant + target company + data type (+ subtype in some cases):
- Provided or Not?
- Perceived Value
- Completeness
- Understandability
- Accuracy
- Useability
- Usefulness
- Meaningfulness
- Feelings about data (general, and company-specific)
- General questions (general, and company-specific)

- Best and Worst Companies (taking into account provided, completeness, understandability, accuracy, usability, usefulness)
- Sankey analysis of participant journeys

2. Text File Processing (Splitting & Recombination)

The researchers then moved on to prepare for the fully qualitative analysis. All interview audio was auto-transcribed using Zoom and Google Recorder, and then the generated text files were cleaned. Cleaning consisted of listening to sections of audio where transcription seemed inaccurate and correcting the transcripts. Due to the volume of data this cleaning was not done for all texts, only where ambiguity or typos meant it was needed for accurate coding and for quotes. Some anonymisation of source texts was also carried out at this stage and later, with a particular focus on quotes included in the chapter. The researchers used this data preparation stage as an initial means of (re)familiarising with the dataset. With reference to the structured interview schedules, the initial 33 text transcripts were split up by participant, company and topic using the labelling scheme outlined in 'Text File Labelling Strategy' below.

At the end of this process, roughly 100 'pieces' had been identified for each participant (slightly more for P11 whose interview 1 covered a broader scope and considerably less for P9 who only did interview 1).

3. Categorisation into CSVs

The pieces from stage 1 were then recombined, across all participants, into 233 source files. These 233 source files were then further grouped into 6 topics areas. (The aim of the analysis was to identify common opinions and ideas around different topics, not to explore individual participant journeys end-to-end). The six topic areas were:

1. **POWER** – discussions and scoring around the power of data holding companies
2. **TRUST** – discussions and scoring around participants' subjective trust in data holding companies
3. **LIFE** – life sketching and annotation discussions, and 'digital life' questioning
4. **HOPES & USES** – discussions around motivations, expectations, goals and hopes, and imagined uses of data
5. **COMPANY-SPECIFIC** – (repeated once per target company per participant) – all discussions around the data return from a particular company
6. **GENERAL** – all non-company specific discussions not captured elsewhere

This produced too many files for import into Quirkos Cloud, so once organised by topic, these six groups of files were further combined into 11 General files and 46 Company-Specific, files (with **Life** and **General** going into the General files and everything else going into **Company-Specific**). This gave 57 organised CSV files ready for use in the first coding phase.

4. Inductive Coding

The majority of the analysis took place with the use of Quirkos Cloud (Daniel Turner, 2014), a computer-assisted qualitative data analysis software (CAQDAS) package that allows for collaborative analysis by more than one researcher. The 57 files from stage 3 were imported into Quirkos Cloud, with each having a unique number. The sources in Quirkos were labelled by Participant, Company and Topic for easy search and retrieval. The researchers then collaboratively coded sections of the interview transcripts to develop and ensure a consistent approach, based on established techniques (Huberman and Miles, 2002; Braun and Clarke, 2006). Codes were identified inductively and not according to a fixed or predetermined set. Once a baseline codeset and strategy had been established, they each coded sections of interviews in parallel, regularly regrouping to discuss generated codes and any new questions or challenges arising. At first, these codes were created in an unstructured/flat state with only occasional clustering on the Quirkos interface. Due to the volume of data, not every piece of every transcript was coded, however care was taken to ensure a representative sample of views from across the participant pool was included. These were clustered into loose code-topic areas, an example is shown in the following screenshot taken approximately 6 weeks into coding:

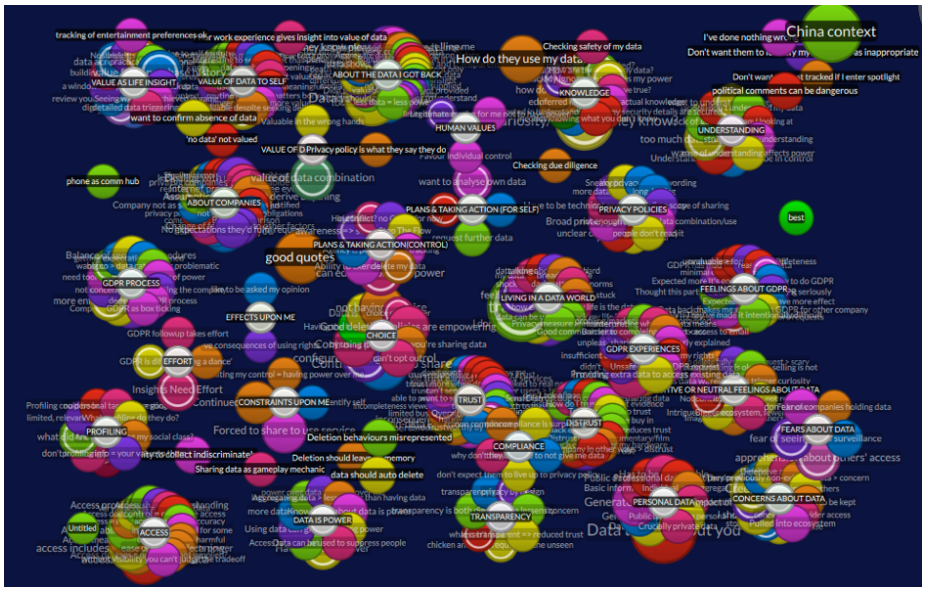

Figure 4: Figure ARI.3: Screenshot from Quirkos During Coding Process

5. Reductive Cycles

As more codes were identified and structures and commonalities between them were formed, existing codes were merged or absorbed into one another and grouped together in small clusters. The researchers regularly met to discuss each other's codes according to their context and occasionally amended wording or

merged concepts that were labelled differently but semantically equivalent. All codes were checked and agreed between these two researchers. Over time, the codes were iteratively structured and restructured, creating top-level thematic clusters around different research questions that held multiple layers of related codes. These clusters were then summarised with a short sentence or paragraph of text, allowing summaries to be produced at different levels of hierarchy. These summaries were kept in the Description fields of codes in Quirkos and also in external structured text-based documents. These can be seen in the following screenshot, taken 5 months into coding:

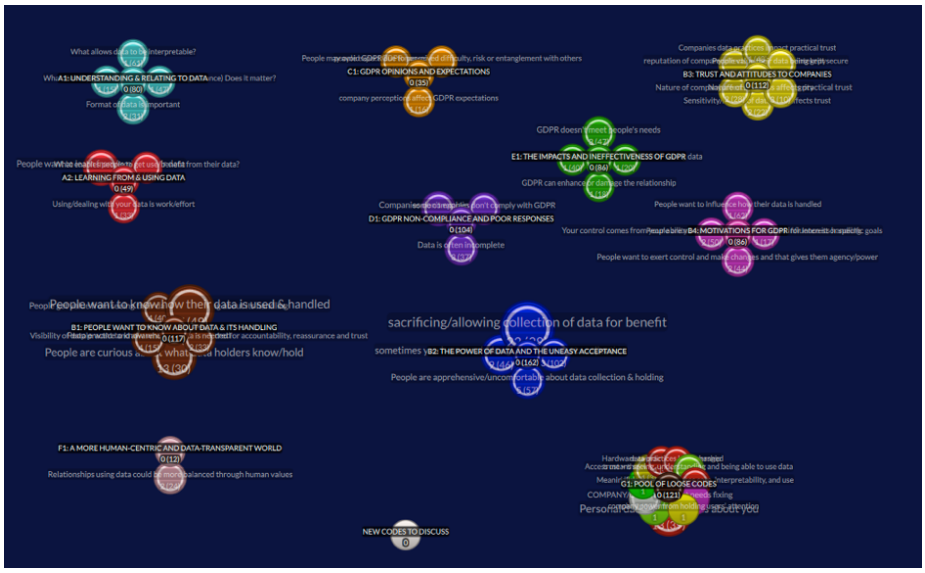

Figure 5: Figure ARI.4: Screenshot from Quirkos at End of Coding Process

The above-pictured structure of the coded corpus at the end of the Quirkos Cloud phase was as follows:

- A (129 codes): What do people/need want from their data and how do they feel about data?
  - A1 (80 codes): **Understanding and Relating to Data**: People want to understand and need to be able to relate to data.
  - A2 (49 codes): **Learning from and Using Data**: People want to learn more from and do more with their data.
- B (279 codes): What do people feel about the data-centric world?
  - B1 (117 codes): **People want to know about data and its handling**: People want to know what data exists and how it is handled, so they can understand what companies do to hold them to account, and inform their choices/trust.
  - B2 (162 codes): **The Power of Data and the Uneasy Acceptance**: People accept a certain amount of data collection and use but are

apprehensive and sometimes feel they have no choice, because data holding is a form of power.
  - B3 (112 codes): **Trust and Attitudes to Companies**: Trust placed in companies is influenced by both the nature and reputation of the company, as well as what data they hold and how that data is kept and handled.
  - B4 (86 codes): **Motivations for GDPR**: People want to make use of their data and influence how it is handled and see GDPR as having the potential to help them achieve this.
- C (35 codes): What do people think about GDPR?
- C1 (35 codes) **GDPR Opinions and Expectations**: People's expectations for GDPR are affected by their perception of the company and its perceived difficulty, risk and entanglement; people expect non-compliance.
- D: What is the experience of GDPR as a means to gain awareness of and access to useable and understandable data?
- D1 (104 codes) **GDPR Non-Compliance and Poor Responses**: The data returned from GDPR is often incomplete, hard to deal with, lacking explanation, or poorly formatted. Many companies are not complying.
- E: What is the experience of GDPR as a means to gain influence and achieve goals with data/What is the practical impact of GDPR?
- E1 (86 codes) **The Impacts and Ineffectiveness of GDPR**: People's interest in GDPR comes from curiosity to exert their rights or from specific questions about data handling or data use goals. GDPR rarely delivers upon on any of their goals but it does change people's outlook and affects the relationship with the data holder.
- F: How should the world change or be different?
- F1 (12 codes) **A more human-centric and data-transparent world**: People want companies to provide greater transparency and data control/agency and act in a more human manner so they can trust them.
- G: Loose/ungrouped codes (121 codes)

Total codes = 645.

6. Theme Identification & Quote Extraction

Having produced the structure above as a reduced representation of *'what the codes say'* that the participants think, the researchers used outlining tool Workflowy (Turitzin and Patel, 2010) to develop the arguments and primary narrative of the chapter into a structured three-theme-based summary of the most important items from these findings. The code hierarchy was used as source material to populate the three key themes with illustrative quotes and observed findings. An example from later in this process (around 8-9 months since Stage 1 began) is shown in the screenshot below:

The themes are broken down in detail in 5.4 and can be summarised as:

1. **Insufficient Transparency**: Organisations appear evasive over data when responding to GDPR, leaving people "in the dark" even after making

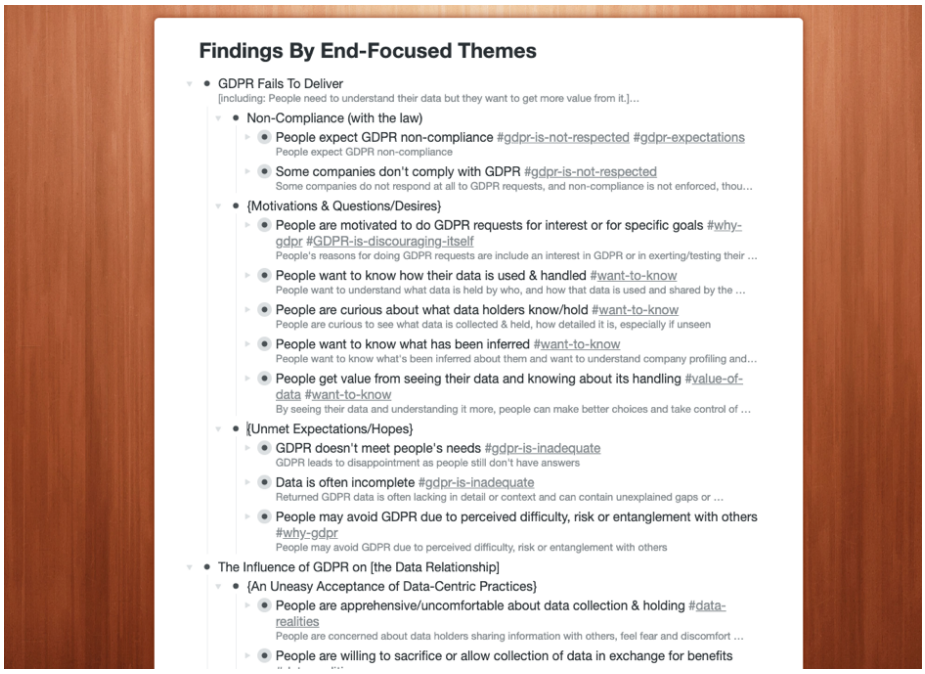

Figure 6: Figure ARI.5: Screenshot from Workflowy During Theme Construction

GDPR requests.

2. **Confusing Data**: When presented with their data, people struggle to understand it and relate it to their lives and are not able to make use of it.
3. **Fragile Relationships**: Companies' data practices, and in particular their privacy policies and GDPR response handling, can be impactful to customer relationships, carrying a risk of damaging trust but also the potential to improve relations.

In all, the process from commencing data analysis to writing up thematic findings in the chapter took over 200 person-hours over a 9-month period from January to September 2020.

Text File Labelling Strategy used in Stage 2

In stage 2, text files were initially broken down into small pieces and labelled according to the following strategy:

Interview 1 (Sensitisation / Poster Display Chat)

Break into 5 parts:

- `Comp` - list of companies
- `Type` - types of data
- `DoWt` - potential uses of data ['what would you do with the data?']

- `GDPR` - GDPR
- `Motv` - motivation for taking part

Interview 1 (Main Sketch Interview)

Break down as follows:

- `SktR` - review of previous sketch interview from prior study [p11 only]
- `DPer` - definition of personal data
- `DAcc` - definition of access to data
- `DCon` - definition of control of data
- `DPow` - definition of power
- `Sket` - sketching
- `Anno` - annotation
- `SelC` - company selection
- `XXXX` - per company [use first four letters of company]
- `Powr` - power
- `Hope` – hopes
- `Uses` – uses
- `Wrap` - [Wrap up]/What happens next

Format: `NN-pXX-iX-[Comp/Type/Uses/GDPR/Motv]-[company first three letters].txt`

e.g. `01-p01-i1-Comp.txt` or `02-p01-i1-Powr-Face.txt`

Interview 2

Break down as follows:

- `XXXX` - per company [use first four letters of company name]
  - `Priv` - viewing privacy policy
  - `Powr` - power
  - `HopU` - hopes & uses
  - `Trst` - trust [p10 & p11]
  - `Pow2` - end power
  - `Trs2` - end trust
  - `Hop2` - end hopes and uses

Format: `NN-pXX-iX-[....]-[company first three letters].txt`

e.g. `01-p01-i2-Priv-Goog.txt`

Interview 3

Break down as follows:

- [intro & consent] - no need to transcribe/code
- `XXXX` - per company [use first four letters of company name]
  - `Powr` - power rating
  - `Trst` - trust rating
  - `RPow` - retro power

- `RTrs` - retro trust
- `Hope` - hope (for company) and uses (how well have hopes been met / how practical are the envisaged data uses
- `Data` - Overall data overview
- `Prov` - Data provided by you
- `Indr` - Data indirectly / automatically collected
- `Derv` - Data derived about you
- `Othr` - Data from other sources
- `Meta` – Metadata
- `GenQ` - general questions about this company
- `Pow2` - end power
- `Trs2` - end trust
- `Next` - what next for this company specifically
- `Genr` - General topics
- `Hope` - Hope (general)
- `Wrap` - Wrap up questions / the future

Format: `NN-pXX-iX-[....]-[company first three letters].txt`

e.g. `01-p01-i3-Cred-Indr.txt` or `02-p01-i3-Genr-Wrap.txt`

Additional 7: Best and Worst Companies for GDPR Handling

The quality and coverage datapoints described in 5.3.3 also allowed insights about which service providers were strongest or weakest in each category, and overall, to be drawn. This was done by tallying the 'Yes' responses for each category and overall, then dividing by the number of times that provider was selected, to avoid inflating scores for popular companies. The outcome of this analysis is shown in Table ARI.2 below. The companies that fared worst overall were those that did not return any data at all in response to a GDPR request (Sainsbury's, Freeprints, Tyne Tunnels, LinkedIn, Huawei, Bumble, LNER). It should be noted that Sainsbury's and Huawei *did* respond, claiming to hold no data for the requesting participant, though participants found this implausible, which indicates either a problem with compliance, explanation or trust. The other named companies here did not respond at all, despite at least two follow-up emails being sent to them, and despite in some cases having initially acknowledged and promised to satisfy the request.

Companies producing responses with good coverage and good quality included Niantic, Nectar and Sunderland AFC as well as to a lesser extent Natural Cycles, Revolut, Spotify, Tesco and Amazon. Facebook and Google fared well for the breadth of data returned (due in part to their download dashboards), though the quality of Google's data was found lacking across multiple categories. Last.fm (owned by CBS) fared poorly overall due to poor category coverage, despite the limited data that it did return being of high quality.

Category / Metric	Best Companies	Worst Companies
Availablity of Data / Breadth of Data Returned	Nectar, Niantic, Sunderland AFC, Natural Cycles, *Facebook, Google, Spotify, Revolut*	Sainsbury's, Freeprints, Tyne Tunnels, LinkedIn, Huawei, Bumble, LNER, Nexus, *Three, Philips Hue (Signify), Check My File*
Completeness of Returned Data	Niantic, *Nectar, Sunderland AFC*	Sainsbury's, Tyne Tunnels, Freeprints, Nexus, LinkedIn, Huawei, Revolut, Bumble, LNER, last.fm (CBS), *Google, Tesco*
Accuracy of Returned Data	Sunderland AFC, Niantic, *Tesco, Nectar, Amazon, Natural Cycles*	Direct Line, last.fm (CBS), *Google*
Understandability of Returned Data	Nectar, Spotify, Sunderland AFC, Niantic, Apple, *last.fm (CBS)*	AirBNB, Virgin Media, *Google, Instagram, Tesco*
Meaningfulness of Returned Data	Niantic, *Spotify, Sunderland AFC, Natural Cycles, last.fm (CBS)*	AirBNB, Credit Karma, Philips Hue (Signify), Direct Line
Usability of Returned Data	Amazon, last.fm (CBS), *Facebook*	AirBNB, Credit Karma, Virgin Media, Sunderland AFC, Huawei, Three, *Google*
Usefulness of Returned Data	Amazon, Facebook, Virgin Media, Spotify, Revolut, Niantic, last.fm (CBS)	AirBNB, Credit Karma, Nectar, Direct Line, Three, *Google*
OVERALL [a]	**Niantic, *Sunderland AFC, Facebook, Spotify***	**Sainsbury's, Freeprints, Tyne Tunnels, LinkedIn, Huawei, Bumble, LNER, *last.fm (CBS), Philips Hue (Signify), Nexus***

[a] Companies were ranked according to total number of all responses in that category for this company that were "Yes" .
[b] Company names in normal text are best/worst; names in italics are second best/second worst.

[PRODUCTION TODO replace table with text]

Table: Table ARI.2 - Best and Worst Data Holders for GDPR, according to Participants' Judgementsa

Additional 8: BBC R&D's Cornmarket Project

I took a three-month sabbatical from my PhD in the summer of 2020. I was remotely embedded within a full-time research internship at **BBC R&D** - the British Broadcasting Corporation (BBC)'s Research and Development (R&D) department (British Broadcasting Corporation, 1997), working with specialists, designers, researchers and developers on an exploratory research project codenamed *Cornmarket*. I continued this involvement as a part-time research consultant and critical friend for a further 5 months after the conclusion of the initial three-month placement.

As part of its Royal Charter, one of the BBC's lesser known obligations is to maintain a centre of excellence for research and development in broadcasting and electronic media, and to this end it employs over 200 researchers in its R&D department looking at everything from AV engineering and production tools to new forms of media, virtual reality, digital wellbeing and human data interaction (British Broadcasting Corporation, 1997). The Cornmarket project, launched in 2019, is a BBC-internal human-data interaction research project which explores a possible role for the BBC as it moves beyond broadcast television, using its public service responsibility to guide citizens to a position of empowerment within today's digital landscape - encompassing not just entertainment but health, finance and self-identity. Due to its unique funding from UK-wide

TV licensing and its duties to not only entertain but to inform and educate the general public, the BBC is uniquely placed to take a more human-centred approach than commercial innovators in this space as it needs only to deliver value, not profit. The project is exploring the use of Solid (Berners-Lee, 2022) technology to build a working Personal Data Store (PDS) prototype [2.3.4] while also developing, iterating and trialling user interface designs and conducting participatory research interviews and activities all to explore what for a BBC PDS might take and what features its potential users might value.

The proposed BBC Cornmarket product, internally called *My PDS*, would allow people to populate a PDS with personal data from APIs and data downloads from a variety of services including BBC iPlayer, Netflix, All4, Spotify, Instagram, Strava, Apple Health, banks and finance companies, as well as social media companies such as Facebook, LinkedIn and Twitter, and then to use these combined data sources to create personal *profiles* for Health, Finance, Media (i.e. entertainment) and Core, within which various data insights, visualisations, capabilities would be delivered. One feature the work explores in depth as potentially valuable to users is the ability to include and exclude certain datapoints from the imported viewing history data in order to present a more accurate, curated view of oneself that could then be fed back to other applications such as BBC Sounds to give better content recommendations.

With a cross-disciplinary team of around 20 people including architects, developers, user experience designers, product designers, innovators, participatory researchers and marketers, and funding to outsource public engagement research to agencies, this project represents a significant player in the emerging personal data economy [2.3.4]. As such the Cornmarket project is a fertile ground in which to learn more from practitioners in the PDE space and to test the learnings of this thesis in practice while also finding deeper insights in response to my research questions - in particular RQ3 which is concerned with the building of more human-centric personal data interfaces in practice.

Much of the work I did during this extended internship can be seen in the designs within 7.4.3, as well as the research report I wrote (Bowyer, 2020a) and internship writeup (Bowyer, 2020b). My work with the Cornmarket project can be seen as the concluding part of one of several action research cycles within my PhD [3.2.2].

A number of articles relating to the Cornmarket project have been published:

- https://www.bbc.co.uk/rd/blog/2021-09-personal-data-store-research
- https://paper.dropbox.com/doc/Building-trusted-data-services-and-capabilities-Us49Ek0nex7yClKughPN4
- https://www.wired.co.uk/article/bbc-data-personalisation
- https://www.theregister.com/2021/10/04/column_data_privacy/
- https://www.thetimes.co.uk/article/bbc-and-sir-tim-berners-lee-app-mines-netflix-data-to-find-shows-viewers-like-lxp002gg8
- https://www.ibc.org/download?ac=18659

- https://www.telegraph.co.uk/business/2022/06/09/bbc-wages-war-online-echo-chambers-unbiased-tech/
- https://parliamentlive.tv/event/index/7d249bcf-78e9-447b-907c-81df72b87542?in=15:01:35

Additional 9: Hestia.ai, and Sitra's *digipower* Project

Following the conclusion of the funded period of my PhD, I took up a near-full-time position as Project Leader and Personal Data Coach at **Hestia.ai** (Dehaye, 2019)), a startup based in Geneva, Switzerland. Hestia.ai is a company conducting research, developing technologies, and delivering training, in the emergent MyData/PDE space [2.3.4]. In essence, the company's mission is to help individuals and especially collectives to more easily obtain and understand data held about them, and to help them visualise, aggregate and make use of that data. It is an example of a **data access and understanding services** company as described in 7.4.5.

I was specifically hired to co-lead the *digipower* project (Härkönen and Vänskä, 2021), for Hestia.ai's client, **Sitra** (Sitra, 1967). Sitra is a non-profit organisation in Finland, funded by the Finnish Parliament and accountable to the Finnish people. The goal of the digipower project was to guide 15 European politicians, civil servants and journalists, through the process of obtaining and exploring their own data. The participants were high-profile VIPs, including the former Prime Minister of Finland and former European Commission Vice President, Jyrki Katainen. The goal was to empower those individuals to better understand the workings of the data economy, so that they might be able to influence others and effect change. One of Sitra's goals is to establish a fairer data economy (Sitra, 2018). Methodologically, the project drew heavily on my own Case Study Two [Chapter 5], adopting a similar method of guiding individuals through the process of making GDPR requests and scrutinising the returned data; I was employed on the project for this expertise. Where it differs from my own Case Study is that the focus of the research was outward, on the data economy and the practices of service providers, rather than inward, on the lived experience of the participants. Other differences included the building and use of software interfaces to provide participants with data visualisations, the use of TrackerControl software to audit mobile phone apps [Insight 12], and the direct analysis of participants' retrieved personal data by the Hestia.ai research team (whereas my Case Study explicitly avoided handling participants' personal data). The project resulted in three reports:

- Sitra's official project report (Härkönen *et al.*, 2022); and
- Two technical research reports by Hestia.ai:
  - A high-level interpretation of models of power and influence in the data economy (Pidoux *et al.*, 2022); and

– A detailed auditing of provider practices, evidenced by examples from participants' data (Bowyer, Pidoux, *et al.*, 2022).

At the time of publication of this thesis (August 2022), I continue to be employed by Hestia.ai, working on the research, design and development of tools to help collectives [Insight 10] with data, make data easier to understand [6.1.2; 7.2.4], and exploring methods to help people 'hack the seams' of digital platforms and services [7.4.4].

Where the BBC internship has helped me to understand the practicalities of connecting people with their personal data in pursuit of Life Information Utilisation [7.2.3], my work with Hestia.ai has helped me understand the practicalities of how people might acquire greater Personal Data Ecosystem Control [7.2.3]. In this sense, both peripheral activities have been highly complementary to developing an overview of the pursuit of HDR in practice.

Additional 10: DERC's Healthy Eating Web Augmentation Project

As a software developer I have been aware for a long time that one of the biggest challenges in building new data interfaces is to gain programmatic access to the necessary data. As part of the trend towards cloud-based services and data-centric business practices, it has become increasingly difficult to access all of the data held about users by service providers. Application Programming Interfaces (APIs) are a technical means for programmers to access a user's data so that third-party applications may be built using that data. Unfortunately, as a result of commercial incentives to lock users in and keep data trapped (Abiteboul, André and Kaplan, 2015; Bowyer, 2018), much of users' data can no longer be accessed via APIs. While GDPR data portability requests do open up a new option for the use of one's provider-collected data in third-party applications, this is an awkward and time-consuming route for both users and developers. **Web augmentation** provides a third possible technical avenue for obtaining data from online service providers. It relies on the fact that a user's data is loaded to the user's local machine and displayed within their web browser every time a website is used, and therefore it is possible to extract that data from the browser using a browser extension; this as another **seam** that can be hacked - see 7.4.4 and Insight 12. Similarly, once loaded into the browser, a provider's webpage can be modified to display additional data or useful human-centric functionality that the provider failed to provide.

In order to better understand what is and is not possible using this technique, I participated part-time from 2018 to 2020 as the sole software engineer in a DERC (Digital Economy Research Centre) project. This project was using the web augmentation technique to explore how researchers could improve the information given to users of Just Eat, a takeaway food ordering platform in the UK. The theoretical basis for this research was published in (Goffe *et al.*,

2021, 2022). While this particular use case does not concern personal data, the technology and techniques being used by the project to exploit the browser seam were considered highly relevant to the exploration of HDR-improving possibilities, and the goals of the research project were also human-centric, and consistent with this thesis's research goals - tackling the hegemony of service providers in order to better serve individual needs.

Additional 11: Special Attribution Note for Chapter 7

This is a note about the attribution of insights within Chapter 7, as the ideas originate quite differently than from the rest of the thesis.

This thesis is my own work. All ideas in Chapter 7 are original. Some of the specific details, theories and ideas presented in Chapter 7 arose or were developed or augmented through my close collaboration, discussion and ideation with other researchers both alongside and prior to the PhD timeframe, including:

- Jasmine Cox, Suzanne Clarke, Tim Broom, Rhianne Jones, Alex Ballantyne and others at BBC R&D;
- Paul-Olivier Dehaye, Jessica Pidoux, Francois Quellec at Hestia.ai;
- Stuart Wheater of Arjuna Technologies and Kyle Montague of Open Lab during the SILVER project;
- Louis Goffe of Open Lab on the DERC Healthy Eating project;
- earlier innovation work with Alistair Croll at Rednod, Montréal, Canada (circa 2011); and
- earlier innovation work with Megan Beynon and Dean Upton at IBM Hursley, UK (circa 2006).

Due to these collaborations and the ongoing and parallel nature of many of these projects to my PhD research, it is impossible to precisely delineate the origin of each idea or insight. In practice, ideas from my developing thesis and own thinking informed the projects' trajectories and thinking, and vice-versa. These ideas would not have emerged in this form without my participation, so they are not the sole intellectual property of others, but equally I would not have reached the same conclusions alone, so the ideas are not solely my own either. All diagrams and illustrations were produced by me, except where specified, and the overall synthesis and framing presented in this chapter is my own original work. Where this chapter includes material from the four peripheral projects [7.1.2], that material is either already public, or permission has been obtained from the corresponding individuals or project teams.

Bibliography

Abiteboul, S., André, B. and Kaplan, D. (2015) *Managing your digital life with a Personal information management system.* 5. ACM, pp. 32–35. doi: 10.1145/2670528.

Berners-Lee, T. (2022) 'Solid: Sir tim berners-lee's vision of a vibrant web for all'. Inrupt. Available at: https://inrupt.com/solid/.

Bowyer, A. (2018) 'Free Data Interfaces: Taking Human- Data Interaction to the Next Level', *CHI Workshops 2018*. Available at: https://eprints.ncl.ac.uk/273825.

Bowyer, A. *et al.* (2018) 'Understanding the Family Perspective on the Storage, Sharing and Handling of Family Civic Data', in *Conference on human factors in computing systems - proceedings*. New York, New York, USA: ACM Press, pp. 1–13. doi: 10.1145/3173574.3173710.

Bowyer, A. (2020a) 'Design research for cornmarket PDS, recommender & associated permissions: Report by alex bowyer (BBC research intern/open lab PhD)'. Available at: https://bit.ly/bbc-pds-research-bowyer.

Bowyer, A. (2020b) 'Designing personal data interfaces - a multi-disciplinary challenge'. Available at: https://bit.ly/bbc-internship-alex-bowyer (Accessed: 18 August 2022).

Bowyer, A., Pidoux, J., *et al.* (2022) *Digipower technical reports: Auditing the data economy through personal data access*. doi: 10.5281/zenodo.6554177.

Bowyer, A., Holt, J., *et al.* (2022) 'Human-GDPR interaction : Practical experiences of accessing personal data', *CHI '22*.

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*. Taylor & Francis, 3(2), pp. 77–101. doi: 10.1191/1478088706qp063oa.

British Broadcasting Corporation (1997) 'Our purpose'. Available at: https://www.bbc.co.uk/rd/about/our-purpose (Accessed: 18 August 2022).

Daniel Turner (2014) 'Quirkos cloud'. Available at: https://www.quirkos.com/learn-qualitative/features.html.

Dehaye, P.-O. (2019) 'Hestia.ai: About us'. Available at: https://hestia.ai/en/about/.

Goffe, L. *et al.* (2021) 'Appetite for disruption: Designing human-centred augmentations to an online food ordering platform', in *34th british HCI conference*, pp. 155–167.

Goffe, L. *et al.* (2022) 'Web augmentation for well-being: The human-centred design of a takeaway food ordering digital platform', *Interacting with Computers*.

Härkönen, T. *et al.* (2022) *Tracking digipower: How data can be used for influencing decision-makers and steering the world*. Sitra. Available at: https://www.sitra.fi/en/publications/tracking-digipower/.

Härkönen, T. and Vänskä, R. (2021). Sitra. Available at: https://www.sitra.fi/en/projects/digipower-investigation/#what-is-it-about.

Hoofnagle, C. J., Sloot, B. van der and Borgesius, F. Z. (2019) 'The European Union general data protection regulation: What it is and what it means', *Information and Communications Technology Law*. Taylor & Francis, 28(1), pp.

65–98. doi: 10.1080/13600834.2019.1573501.
Huberman, M. and Miles, M. B. (2002) *The qualitative researcher's companion*. Sage.
Information Commissioner's Office (2014) 'Data controllers and data processors: what the difference is and what the governance implications are', p. 20. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/introduction-to-data-protection/some-basic-concepts/.
Moraveji, N. *et al.* (2007) 'Comicboarding: Using comics as proxies for participatory design with children', in *Conference on Human Factors in Computing Systems - Proceedings*. ACM, pp. 1371–1374. doi: 10.1145/1240624.1240832.
Pidoux, J. *et al.* (2022) *Digipower technical reports: Understanding influence and power in the data economy*. doi: 10.5281/zenodo.6554155.
Sitra (1967). Available at: https://www.sitra.fi/en/topics/strategy-2/#what-is-sitra (Accessed: 18 August 2022).
Sitra (2018) 'Sitra's fair data economy theme: What is it about?' Available at: https://www.sitra.fi/en/themes/fair-data-economy/#what-is-it-about (Accessed: 18 August 2022).
Spinuzzi, C. (2005) 'The methodology of participatory design', *Technical Communication*. Society for Technical Communication, 52, pp. 163–174.
The European Parliament and the Council of the European Union (2016) 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data', pp. 16–32. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=ES.
Turitzin, M. and Patel, J. (2010) 'Workflowy'. Available at: https://www.workflowy.com/features/.