# Appendix 3 – Coding and Analysis Approach

All coding was carried out by the first and second authors, 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 timecoded text sections from text files into cells of 6-topic spreadsheet, then generation of CSV files for importing into Quirkos Cloud [3]
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 [4] 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
  + Usability
  + 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 paper. 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 [3], 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. Authors 1 and 2 then collaboratively coded sections of the interview transcripts to develop and ensure a consistent approach, based on established techniques [1,2]. 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:

Background pattern

Description automatically generated

Figure 7: 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 authors 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 authors. 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:

Timeline

Description automatically generated

Figure 8: 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 usable 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 first two authors used outlining tool Workflowy [4] to develop the arguments and primary narrative of the paper 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 paper’s 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:

Text

Description automatically generated

Figure 9: Screenshot from Workflowy during Theme construction

The themes are broken down in detail in Section 5 of the paper 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 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 paper 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

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

## Bibliography of this Appendix

[1] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2: 77–101. https://doi.org/10.1191/1478088706qp063oa

[2] Matthew B. Miles, A. M. Huberman, and Johnny. Saldaña. 2014. *Qualitative Data Analysis: A Methods Sourcebook*.

[3] Quirkos Cloud. https://www.quirkos.com/learn-qualitative/features.html

[4] Workflowy. https://www.workflowy.com/features/