Where's My Money?
A Critical Report on Receipt Data Collection & Datafication

Alexander Juby-Rasmussen (AJR), AU694667, 202208686@post.au.dk

5th Semester, Cognitive Science BSc, Arts Faculty, Aarhus University

Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Critical Data Studies, Elective

Curating Data

Characters: 4.800

Process of Collecting, Digitising, and Datafying

Between 15 June 2025 (21:16) and 24 September 2025 (10:32), I collected 27 physical receipts from everyday purchases, primarily at Netto, but also from other shops such as Rema 1000, Magasin, and Faros Cigarer. These receipts served as traces of my household spending and were digitised to support account balancing within my three-person apartment. The aim was to ensure that costs for shared items were distributed fairly across all members of the apartment, while also allowing me to be reimbursed for money I had spent.

It began with **collection** of the paper receipts. Each receipt was photographed and logged in an Excel sheet. For each receipt I recorded a set of variables that were immediately available on the receipt itself: shop name, date, time, number of unique items, amount spent in Danish kroner (DKK), number of unique item discounts, and total VAT (MOMS). Not only did I select these variables, I also looked at: the quality of the receipt (poor, good, excellent), whether there was a separate bank statement, who serviced the transaction (e.g., self-service or a cashier), and the physical length of the receipt measured in centimetres using the observer's (me) perception and a ruler.

After collection, I begun the process of **digitisation** via Excel. Totals were written in base-10 (the standard form used across most countries *('Decimal', 2025)*) and VAT amounts were written in DKK to 2 decimal places; time was standardised (hh:mm:ss); physical length was measured in centimetres to 1 decimal place; and text fields such as shop and server names were normalised where possible. Quality consistency was done by purely by the observer's opinion (mine) and judged on an internal scale which resulted in the 3 choices of: poor, good or excellent.

Finally, I looked at the process of **datafication**. The receipts were not only archived but turned into structured data that could be sorted, filtered, and aggregated. Dates and times allowed ordering, and reimbursement from my flatmates; the "Amount Spent" field became the key to calculating the per-person amount; and the "Unique Item Quantity" variable gave an estimate for the complexity of each purchase the physical length of a receipt, once transcribed, became a number that could be graphed or compared. In short, the receipts were transformed from messy paper traces into a proper dataset.

Critical Reflection on Design Decisions

The choice of variables was primarily driven by what receipts make available. Shop name, date, time, and totals are obvious choices, as they are core variables to account for when accounting. The decision to include "Quality" reflects an awareness that not all traces are equal: poor print quality makes full comprehension and transcription of the content more prone to error. This was also decided on the quality of the paper and the frequency of creases found throughout.

The physical length of receipts was both measured but was included as a potential proxy for the number of items. This creative solution is also problematic: receipts differ in formatting, line spacing, and whether the bank statement is contained or not; so length cannot be taken as an objective indicator. When I receipt length as a proxy for item count, or when I come to input varying splits across all three household members, I am effectively generating synthetic data (Steinhoff, 2024). These values are derived and are not simply found, containing a variety of assumptions that may or may not hold. This reminds us that datafication always involves interpretation and should not be taken at face value.

"Serviced by" was mostly recorded as "Self," which provides little discriminatory power outside the context of the logged data which for the household accounting task is not necessary but may be deemed useful for an external viewer. In Excel, receipts become rows and features columns, enabling analysis but also constraining it. Formulas for either equal splits or exact per-person-split appear objective yet obscure contextual decisions about who benefited. Therefore, spreadsheets encode accountability while hiding the incidents of collection and curation (Dourish, 2017).

By focusing on receipt-visible variables, the dataset is strong for auditability but weak for modelling consumption; VAT, discounts, and length enable analysis but don't resolve reimbursement. Missing data on which flatmate owes me leaves the key question of who owes what unanswered. As datastructuring is an active work: my choices in variables, quality, and reimbursement shaped how receipts became actionable yet obscured the backstage labour (Flyverbom & Murray, 2018). The backstage labour of transcribing, cleaning, and normalising is precisely what makes the dataset work for reimbursement.

Bibliography

Decimal. (2025). In *Wikipedia*.
https://en.wikipedia.org/w/index.php?title=Decimal&oldid=1306818039

Dourish, P. (2017). Spreadsheets and Spreadsheet Events in Organizational Life. In *The Stuff of Bits: An Essay on the Materialities of Information* (pp. 81–104). MIT Press.
https://ieeexplore.ieee.org/document/7932792

Flyverbom, M., & Murray, J. (2018). Datastructuring—Organizing and curating digital traces into action. *Big Data & Society*, *5*(2), 2053951718799114.
https://doi.org/10.1177/2053951718799114

Steinhoff, J. (2024). Toward a political economy of synthetic data: A data-intensive capitalism that is not a surveillance capitalism? *New Media & Society*, *26*(6), 3290–3306.
https://doi.org/10.1177/14614448221099217

Bibliography