Praise for *Between the Spreadsheets*

‘Everyone talks about why data is important, but rarely do people bother to talk about how to make data useful, especially if it is dirty. Susan does just that by giving readers very clear and practical details on how to clean, structure, and classify data in a sustainable way. The lessons are clear and imbued with Susan’s wit and experience, which makes learning this complex topic fun. This should be a foundational text in all procurement training and university programs.’

Dr Elouise Epstein, ***Partner at Kearney and author of ‘Trade Wars, Pandemics, and Chaos’***

‘Only Susan could write a book like this! Dirty data is a problem facing every organisation and Susan takes you step by step through the process for fixing this in a pragmatic, straightforward and simple fashion, but more importantly she brings the subject to life. In this book she perfectly balances being informative with being engaging.’

Caroline Carruthers, ***Chief Executive of Carruthers and Jackson and bestselling author***

‘“Garbage in is garbage out”. The world is inundated with data, everyone wants a piece, but how do we make sure the data is usable? Susan Walsh is a leader in this space and has written a wonderful book all should read on dirty data and classifying it correctly. This is both for the preparer, the user, and everyone who is looking to take part in the data revolution.’

Jordan Morrow, ***the ‘Godfather of Data Literacy’***

‘Susan’s writing style is unique and effective at making the technical data quality topics highly enjoyable to consume. Her brilliant idea of data getting a COAT on is a powerful method for guiding readers through the steps required for protecting your data quality.’

Kate Strachnyi, ***Founder of DATAcated***

‘This book is a page turner. Susan’s unique style keeps you engaged and craving to learn more from the practical steps and examples she so carefully presents in this book. This is a must read for both business and data professionals who want to clean up their data and get more out of it.’

George Firican, ***Founder of LightsOnData and data governance expert***

‘Susan Walsh is a force of nature in the data business. She brightens any LinkedIn feed while broadcasting live from her Data Den in front of pink glittery streamers or posting crazy lip-sync videos. In *Between the Spreadsheets*, Susan’s vivacious humor and unstoppable spirit livens up an otherwise traditionally boring topic – Data Quality. But inaccurate, inconsistent, untrustworthy data remains a serious problem that plagues literally every company. And Susan can help.

Whether you learn from her practical hands-on tips and tricks or her inspirational life story you’re sure to leave with more than you came with (and knowing Susan, probably more than you bargained for!).

So grab a glass of Prosecco, don your Data COAT and slip Between The Spreadsheets to learn how to clean that dirty data of yours, once and for all!’

Scott Taylor, *The Data Whisperer*

Between the Spreadsheets

Between the Spreadsheets

Classifying and Fixing Dirty Data

Susan Walsh

facet publishing

Published by Facet Publishing
7 Ridgmount Street, London WC1E 7AE
www.facetpublishing.co.uk

Facet Publishing is wholly owned by CILIP: the Library and Information Association.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN 978-1-78330-503-2 (paperback)
ISBN 978-1-78330-504-9 (PDF)
ISBN 978-1-78330-523-0 (EPUB)

First published 2021

Text printed on FSC accredited material.

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C013604

Typeset from author's files in 10.5/13pt University Old Style and Myriad Pro by Flagholme Publishing Services.
Printed and made in Great Britain by CPI Group (UK) Ltd, Croydon, CR0 4YY.

Contents

Figures

Tables

Acknowledgements

Firstly, I have to thank my dad, Bill, my biggest supporter, who has helped me through some really tough times without ever wavering in his support or love. And he still continues to give me advice, knowing that I probably won't take it. Love you lots.

Then there's my brother Will. He's the voice of reason sitting on my shoulder asking 'is that wise?' He and his wife Kristianne (and not forgetting Rocco the dog) have been great cheerleaders when things have been tough and I thank you all for that greatly. Love you guys.

And then there's my mum, Joan, who's no longer with us. She would have been so unbelievably proud and the whole of Dundee would have known about this book before it had even been written. But she'd also be very disappointed I've not been married off yet. I hope this makes up for it, Mum, miss you and love you loads.

Next, it has to be Vicki Connor. A straight-talking, no-nonsense coach who metaphorically pulled me off the road at the last minute before a truck of destruction hit. She never took advantage when I was down and did quite the opposite – lifted me up higher than ever before. With her coaching, my mindset and self-belief has changed for the better and I don't think I'd even have considered writing this book without her subconscious ninja mind tricks.

Followed closely behind is Michelle Henty and Helena Jannson West. My bestest friends in the whole world who have peeled me off the floor when things have been down, or off the ceiling when life's been good. Your support has been invaluable and we've kept the Prosecco industry alive in the last few years. I look forward to sharing many more Proseccos with you both.

None of this would have been possible had it not been for a chat with Caroline Carruthers, who believed in me enough to introduce me to my now publisher, Peter Baker at Facet Publishing. You both saw the potential I knew was there and I thank you both for seeing that and embracing the crazy.

Then there are the book reviewers, the ones who gave honest and constructive feedback. Thank you so much Natalie, Kavita, Kirsty, Danielle, Kate, George, Michelle, Adriana and Edward.

To Ferdos, Sarah, Tracey and Tiff, my oldest friends – thank you for being there, always.

Kay and Eric, you've been cheering me right on too, even though you have no idea what I do. The best kind of family.

The legend that is Scott Taylor, the Data Whisperer, can't go without recognition. He welcomed me into his group, showcased me to the world through 'The 12 Days of Data Management', and through him I got to know Kate Strachnyi and George Firican. We've been firm friends ever since, despite the distance. I'm so lucky to know you guys.

To my procurement ladies, in particular Sarah, Sarah and Amanda. Thank you for your support and friendship, you inspire me every day.

Finally, to every single follower and connection on LinkedIn, I thank you. You've embraced my unique style and given me so much fun and joy. Thank you.

I know I've probably forgotten someone, so please use this as your acknowledgement!

Abbreviations

AI	Artificial Intelligence
B2B	Business to Business
B2C	Business to Consumer
CIPS	The Chartered Institute of Procurement & Supply
COAT	Consistent, Organised, Accurate, Trustworthy
CPO	Chief Procurement Officer
DPO	Data Protection Officer
EEA	European Economic Area
EU	European Union
EUR	Euro
GBP	British Pound Sterling
GDPR	The General Data Protection Regulation
GL	General Ledger
HR	Human Resources
HVAC	Heating, Ventilation and Air Conditioning
ICO	Information Commissioners Office
ML	Machine Learning
MRO	Maintenance, Repair and Operations
PO	Purchase Order
RFID	Radio Frequency ID
RPA	Robotic Process Automation
SKU	Stock Keeping Unit
UNSPSC	United Nations Standard Products and Services Code
USA	United States of America
USD	United States Dollar

Introduction

Hello and welcome to *Between the Spreadsheets: Classifying and Fixing Dirty Data*.

This is not your typical data book and that's because I'm not your typical data person. I have a wonderfully unique background that's a mix of corporate and data work, which has brought me to the point where I'm able to share my specialist knowledge with you.

Regardless of whether you are completely intimidated by data, starting out in your data career, a seasoned procurement or data professional, or a decision maker within an organisation, there will be something in here for you.

Dirty data is a problem. In every single organisation, no matter how big or small or where they're located, you will hear talk of data quality issues. What you will rarely hear about is the consequences of this because people or companies don't want to admit their failures. We could be talking about millions of pounds or dollars lost on new technology, weeks or months spent fixing mistakes due to bad data, possible job losses or even worse.

It's not just that. We hear all the time of data scientists spending anything from 40-80% of their time cleaning or wrangling data. Why is this? Well, I believe it's because they are inefficient and inexperienced at it. 'What? But they're data scientists!' I hear you cry. Unfortunately, that doesn't mean anything. Data cleaning is rarely covered in academic studies or other courses; the focus is always on the technical aspects of the role, yet ironically, they can't do any of that without the clean data first.

While data cleaning is one of the most vital parts of the whole process when working with data, it is often overlooked because either there is an assumption that people already know how to do it or it's considered too menial or not important enough to spend time on, or invest in.

It's not just in data science; I've seen this in other areas. I was managing a politics student and asked her to book a meeting and send invites. When it hadn't been done, I asked why and she explained she didn't know how to do it. I was blown away that someone studying politics (don't ask me why the politics thing made a difference, I have no idea) didn't know how to create a meeting invite. My assumptions were very wrong.

Then there was the time when I was in a new role and being trained by someone in his late twenties, at least ten years younger than me. He was using right-click to cut and paste all the data in his spreadsheet, so I showed him the shortcut of ctrl+c to copy and ctrl+v to paste. Can you imagine the amount of time he must have spent right-clicking? All those seconds throughout a day add up over the course of a year, plus it must have been hard on his hand doing all that repetitive clicking.

If you thought that was the story, well you're in for a treat. When we finished that training session, he said to me 'and just click this x in the corner to save'. I am not kidding. I put my hands up to my face in horror as he told me this. My very core was shaking, I had cold sweats and I made him promise never to do that ever again. Can you imagine if he lost his work? That's all hours or days of work lost to the business. Thankfully, we have autosave now, but there are still occasions when this doesn't work properly and you shouldn't just take it for granted.

There is a huge gap between what industry assumes or expects of students and what their skill levels are. I want to raise this as an issue to be aware of and use this book to help. Even *The Guardian* (2021) has reported that there's a sharp fall in the number of people taking IT courses. Why is this? Well, let's face it, data can be a very dry and boring subject. (Except when you talk to me of course. I have a data den with pink glittery streamers and do lip-sync videos on LinkedIn.)

I am everything that you don't expect someone working in data to be. And that is a very good thing. It's also why I have a fun book title: I want non-data people to see the title and consider reading this. I want this book to be seen by young people as an opportunity to have fun with data, to see that you can make of it what you want and be who you want and still do a great job while educating people on how to have better data.

It's not just the young people looking to enter into data that struggle with a skill gap; I often hear stories of graduates getting their first jobs in an organisation and being handed their first 'real' data set and they don't know what to do with it. They have been used to working with sterile data sets in a test environment in academia and they don't know how to deal with a real live working data set because they've never seen one in context before.

As procurement moves toward digitisation, there are data quality issues prohibiting this. It wouldn't be a stretch to assume that a mixture of lack of skill and lack of investment in these areas is at least causing some of these data quality issues.

The Deloitte Global Chief Procurement Officer Survey 2019 found that 60% of Chief Procurement Officers (CPOs) stated poor master data quality, standardisation and governance as the biggest challenge for mastering digital complexity. 57% said that the quality and accessibility of data presented strong barriers to technology adoption. These are all things I will address in this book.

I want to highlight that clean data is an investment, not a cost, and I want to demonstrate exactly how much of an investment it is – you could be saving so much more than just time and money. You don't have to invest in super expensive consultants or some fancy software; it can be done in Excel. It might not always be the most efficient way, but for those with limited budgets, it is a way.

I'm here to tell you that it's really, really, *really* important to have good data quality and that if you or your team are not experienced in this area, get reading! Not only will it help make your job easier, but it could potentially save you bucketloads of time and money and give you peace of mind. Now, isn't that worth it?

I think it's important that from the top to the bottom of an organisation, every single person should understand the impact of dirty data and how to spot it. In this book, I'm going to explain why and show you how.

After reading this, not only will you be able to work with your data more efficiently, but you'll also understand the impact that the work you do with it has and how it affects the rest of the organisation, including your colleagues.

Before we get into the nuts and bolts of this, let's answer what will probably be your first question: who is Susan Walsh, The Classification Guru, and what makes her qualified to talk about this?

Let's start at the beginning of my career. I'd never really known what I wanted to do or felt like I had a calling. I decided to study for a degree in Commerce as I thought it would cover many areas and that I would definitely know what I wanted to do when I graduated. Guess what? I didn't.

I grew up in a place called Broughty Ferry in Dundee, Scotland, where the main career options were retail or call centres and I knew those weren't for me. When I graduated, I got a job as a paint merchandiser and that company moved me down to Guildford in England, where I still live 20 years later. But, as you may have already guessed, that job did not last.

For the remainder of my twenties, I tried out a number of careers that mainly revolved around sales. Sales rep, telesales executive, account manager, national account executive and national account manager. All within large blue-chip companies, working with some well known retailers.

Somewhere along the way, I realised I was doing what I *thought* I should be doing, rather than what I *wanted* to do, even though I still didn't know what that was. Then I had an idea: 'let's open a shop!' It was at a time when working in the corporate world was changing for women. It was less suit and blouse and more smart/casual, pretty dresses and tops. Online retail hadn't really taken off yet and it was harder to source nice clothing, so I decided to open a shop. In Guildford. One of the most expensive towns in the UK for commercial rent.

However, I was confident. I had designed a beautiful shop with lovely clothing at a reasonable price . . . and it bombed. Months and months went by with a few customers here and there. Some days no one came in at all. It was soul-destroying. In my gut, I knew it wouldn't work, but I tried to hang on for as long as possible.

What I could never have planned for was the level of brand snobbery in my town. If it wasn't Chanel or Gucci, they weren't interested. I had people walk by for months before they would be comfortable even stepping foot in my shop. It was so hard. At one point, I had a return of an item after Christmas and I barely had enough money to refund the customer – it was really tough.

'This is all very interesting, but what does this have to do with the book?' I hear you say. Well, the next part is where my data journey began.

I had literally no money. I couldn't even afford to go bankrupt. I had to get a job to save up and pay for it. I was desperate for work and I found an online ad for some data entry work for a Spend Analytics company. Although I had never worked in Procurement or Data, I thought I'd be good at it as I had worked for many corporates and knew what they were buying and how their budgets worked, but more than that, I just needed a job. Little did I know how good I would actually be at that job and, more importantly, how much I'd enjoy it.

I found that classifying data came very naturally to me. I understood the context of how the data was being spent, which in the spend data world is not always what it seems to be. (But I'll get into that later.)

I also had to work with a number of analysts, which opened up a whole new world of different languages and communication that was different from what I was used to in the corporate world. After some time working in that environment, I was able to understand both perspectives and become a translator of sorts between the two worlds.

Five happy years followed, and, as the business grew, so did my responsibilities. By the end, I was managing a team of 14 who I'd recruited, trained and managed myself. Through that, I learned how to train and share my knowledge and processes with my team.

Eventually, I needed to move on and find new challenges. I knew I wanted to do that within a similar role, but since I had not come from a procurement or data background and I didn't code or have the experience to find a job as a data analyst, I didn't know where I could find a similar role elsewhere. I decided to start my second business – The Classification Guru.

I wanted to be different and unique. I wanted to be myself in all its bubbly craziness. I didn't want to just replicate what my previous company and many others had done or be another boring data person. I wanted to focus on the one big area I could see was being neglected – **spend data classification and data quality**.

I knew that spend data classification had always been offered as part of another service, such as dashboards and analytics, or if a new system or software was purchased, but it was never available as a standalone service. The perception in the industry was always around the value of the software and the analytics, not the cleaning and preparation of the data. Having spent many years tidying up messy data, I could see that the true value was in the quality of the data, yet no one seemed to be speaking about it.

I started talking about it. A lot. To anyone who would listen. Without a single connection in the procurement or data world, I went out and networked, exhibited and used social media to spread my message. I quickly discovered that not just procurement spend data quality but all data quality was an issue and that it was, and still is, a struggle to raise its profile within organisations. Let's face it, it's not the most exciting topic and it intimidates a lot of people. It can be seen as menial work with little value, when in fact it drives everything an organisation does, from cost savings and efficiencies to production and planning to detecting fraudulent activity.

I saw an opportunity to make it more fun and interesting. Yes, it's a serious subject with some highly skilled and knowledgeable people working in the industry, but they are the ones who already know and understand the issues that organisations face with data quality. The people who need to be engaged are the ones in the rest of the organisation, the ones working with the data, from the receptionist to the planners and forecasters, sales, marketing and supply chain, to the senior executives using the data to make critical business decisions. It needs to be more interesting to be relatable.

To some, it can seem complex and intimidating. Those people very rarely understand the importance of the accuracy of the data they are working with, how it's used or the consequences of when it goes wrong.

That's why I'm writing this book. Whether you have never been interested in data and find it daunting, you're a seasoned professional who is looking to refine or reaffirm your skills, or you're a senior decision maker who wants to understand the impact of dirty data and the value to your organisation, then this book is for you.

I will take you through my definition of dirty data – what is it and what the consequences are of how this affects the rest of the organisation. I'll also show how you can help ensure data accuracy using my COAT methodology, plus the importance of keeping that data COAT on with regular maintenance and spot-checking.

I'll then take you through how to classify, cleanse and normalise, firstly in Excel, as that's a global tool that is accessible to everyone and can be followed by any reader, regardless of skill level or experience. Screenshots and examples will help take you through the processes. This will be followed by me sharing some of my tips with you on how I classify in a different tool and then I'll explore other classification and cleansing options.

If that's not enough, I'll present to you my dirty data maturity model. Yes, it exists and you can gauge where you are and look at steps on how you can improve your dirty data. Finally, I will share with you some data horror stories - there's nothing like finishing up with some drama and suspense, but with a twist!