
# DDRC WORKSHOPS: EMOTIVE CONCERNS SURROUNDING DATA

PREPARED BY: SUZANNE MCCLURE, DDRC PREPARED: MAY 2024

## ABSTRACT

The findings presented in this report examine DDRC workshop transcripts at the thematic and word level, revealing the overarching concerns of participants. The focus is on the identification of common themes and high frequency words within these categories, as expressed by those attending the workshops. Research findings on three significant themes related to emotive words are presented in this report along with transcript extracts. The report findings identify similar patterns across sectors, resulting in an overview of concerns surrounding data as expressed through emotive language. A brief overview of the DDRC workshops is provided in the first section. Quantitative results are then presented and explained, offering significant insights into the participants' thoughts and opinions surrounding data.

## DDRC WORKSHOP OVERVIEW

DDRC workshops were conducted in England, Scotland, and Wales for six UK sectors. The workshops consist of a 1-minnute challenge for all participants and a 45-minute team model-building activity. Various means were utilised for recruiting organisations to participate in the DDRC workshops such as social media and professional contacts. Recruitment materials included a short video with workshop testimonials; an invitation letter from Principal Research, Professor Simeon Yates; and an infographic explaining the workshop format located in Appendix 1. The DDRC workshops are 90 minutes and follow the format presented in Table 1.

| Minutes | Activity |
|---|---|
| 5 | Collection of consent forms and introduction to LSP® methodology |
| 10 | 1-minute Lego challenge warm-up exercise |
| 45 | 45-Group challenge and building of model |
| 30 | Group sharing of their model (i.e. Research dataset) |

**Table 1: Workshop Format**

For the 45-minute group challenge, each team is provided with a large box of Lego bricks and a smaller container with Lego body components; pieces deemed metaphorical such as flags and flames; characters such as fish, crabs, and ducks; and various ladders and connectors. Eight workshop challenges were designed to illicit attitudes towards data for the group activity and these are listed in Table 2.

| 45-Minute Group Lego Challenges |
|---|
| Working as a team, build a mode of data privacy concerns in your organisation. |
| As a team, build what comes to mind when thinking about the value of data in achieving organisational success. |
| As a team, build what comes to mind for improving digital technology management in your organisation(s). |
| Working together, build what comes to mind for raising awareness of data risks in your organisation(s). |
| As a team, build a model of how opportunities in data management can be pursued. |
| Working together, build what comes to mind when you think of a data-driven organisational culture. |
| Working together, build a model of possible barriers to accessing data needed for organisation success. |
| Working together, build a model that expresses concern for the confidentiality and the integrity of organisational data. |

**Table 2: 45-Minute Group Challenges**

As the group challenges were not relevant to workshops conducted for Small and Medium Enterprises (SME) and pensioners, two additional challenges were created for these groups: *Build a model of the information life cycle of personal data in your business, assessing risks to individual privacy and measures that might mitigate these issues* and *Build a model of the information life cycle of personal data in your*

*business and identify unforeseen or unintended uses of data.* These organisations and individuals are represented together in the following research findings as the sector *Enterprise*.

## CONFIRMATION OF WORKSHOP OBJECTIVES

The main objective of the DDRC workshops was to investigate issues pertaining to participants' attitudes towards data. To confirm discussions centred around relevant topics, Sketch Engine (Kilgarriff et al. 2014) was employed to identify language choices that differ from typical spoken British English by comparing the workshop datasets to the British National Corpus (BNC) of Spoken English (2014). The BNC Spoken English corpus contains transcribed recordings made by British English speakers residing in the United Kingdom and is largely comprised of spontaneous spoken English. The analysis provides lexical evidence of language that is unique and statistically significant in the workshop discussions as compared to everyday spoken English.

The 25 highest ranking marked content words in the DDRC workshop transcripts compared to the BNC Spoken English corpus are: trustworthy, data, accessible, firewall, secure, organisation(al), confidentiality, phishing, classification, breach, VPN, GDPR, access, governance, silo, Sharepoint, dataset, stakeholder, priority, untrustworthy, CRM (customer relationship management), privacy, external, integrity and vulnerability.

The top 25 Multiword Expressions (MWE) in the DDRC workshop transcripts are: low priority, secure datum, priority datum, privacy concern, organisational success, type of data, accessing datum, data warehouse, personal datum, lot of data, trustworthy datum, good datum, data risk, data breach, red flag, organisational datum, digital technology, bad datum, open source, critical datum, low priority datum, data privacy, technology management, digital technology management, and different type of data.

These results show with a 99.99% confidence level that the language used in the DDRC workshops differs significantly from everyday spoken British English by including words and MWEs that pertain to data, technology, and security.

## RESEARCH DATASET & QUANTITATIVE SOFTWARE

The DDRC facilitated 33 workshops between June and November 2023. A listing of participating organisations by sector and a word count based on workshop transcripts is contained in Appendix 2. The

word counts for edited workshop transcripts have been determined by MS Word; word count is often necessary for statistical measurements in quantitative research. Edited transcript files have been created by removing software-generated speaker information and standard text, and non-anonymous references. Speaker time stamps remain in the transcripts for qualitative research purposes. The edited DDRC workshop transcript dataset is the focus of this report and contains 125,079 words. The DDRC workshops are divided into six sectors and Figure 1 illustrates the dataset contribution by sector based on word count.

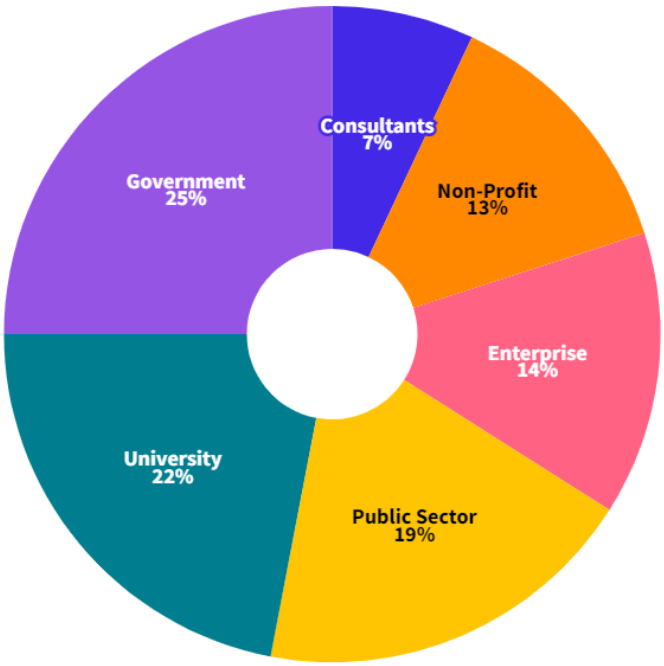



**Figure 1: Sector Composition Based on Word Count**

The edited transcript dataset was processed by Wmatrix (Rayson 2009) for quantitative analysis. The application is utilised in this report to identify prominent themes and high-frequency words within each theme. Wmatrix functions by identifying data that is critical for textual analysis; it does not disregard grammar and therefore makes lexical distinctions. An example of this feature is that the application would identify *dance* in *She went to the dance* as a noun but in *She danced alone* the word would be tagged as a verb. Every word in a dataset is assigned multiple tags for a word's thematic concept (also known as semantic domain). In this analysis, only the first semantic domain tag identified by Wmatrix is used, offering a 91% accuracy rate for a given word in context (Rayson 2019).

## PROMINENT WORKSHOP THEMES

Thematic concepts, or semantic domains, are identified by Wmatrix at the word level and are indicative of the "aboutness" of a dataset. The words within a classification relate to the same perceptual notion and include synonyms, antonyms, hypernyms, and hyponyms. These marked concepts can contribute to the identification of thematic concerns within a dataset. For this report, key themes in the workshop transcript dataset reveal repeated and lexically significant language employed by the DDRC workshop participants.

Examining the high frequency semantic domains shows key concerns as expressed by the language spoken in the workshops. Numerous overarching categories were identified by Wmatrix and the focus of what follows is the three most common emotive domains: *Angry, Discontent,* and *Worry*. The importance of these themes for research is that the high frequency words in each category can be used to easily locate repetitive language patterns, signally an array of concerns surrounding data. Having identified these themes through quantitative analysis, a qualitative examination of the text surrounding these words (co-text) can reveal significant insights into the DDRC workshop participants' views. In the following three sections, each of the significant emotive themes are explained and excerpts from the workshop transcripts are presented. The calculation for distribution by sector has been normalised by dividing the total words assigned to a theme by the total words for that sector.

## ANGRY

By examining the transcript research dataset for the phrases and sentences surrounding high frequency words in the emotive domain of *Angry*, a narrative emerges of DDRC workshop participants' preoccupations with the emotion of anger and perceived risks. The marked words in descending order of frequency in this category include but are not limited to: *threat*, *attack*, *angry*, *annoy*, and *malicious*. Figure 2 shows the contribution by sector for words assigned to the domain of *Angry*.

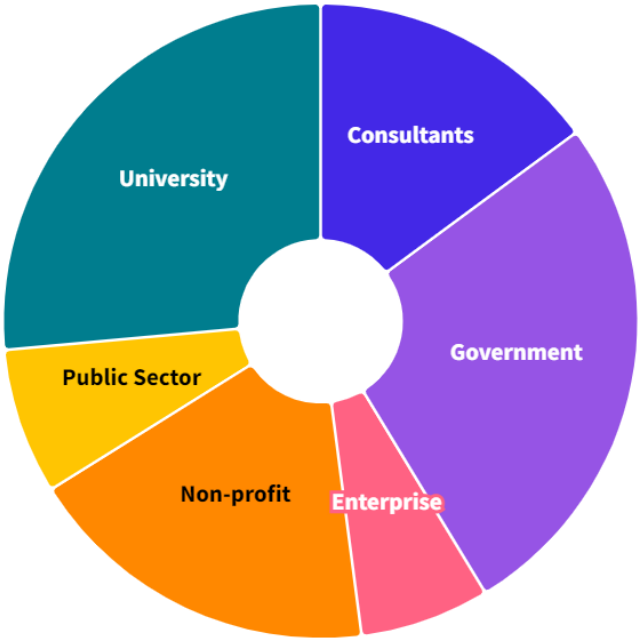



**Figure 2: Angry Theme by Sector**

There is an unequal distribution by sector which represents dissimilar language use; the Enterprise sector held the fewest discussions in relation to the emotion of anger and potential risks. The concerns expressed in this category were discussed by participants in various organisational roles and sectors. The commonality of their stories is illustrated below in selected workshop transcript extracts (key words in bold).

> *Accessibility of data and accurate documentation*: I'm searching through a notes and then another notes, trying to identify is this the data or I want can actually use this data. It's just and then it's not all stored in the same place. So I mean, this is like the ideal world where everything is where it's supposed to be. But that doesn't often always happen, which is a bit **annoying**.
>
> *Cloud data and external threats*: And up here we have a black hat hacker. He's wearing a black hat who and like the scale of the tower symbolises how like you need to take particular account of like there may be **malicious** people out there and then money burning that's just because the cloud costs a lot of money.
>
> *Accessibility of data***:** So yea, storage and transport is not always straightforward, and there's a very **angry**, very **angry** that data not being able to access it.
>
> *Accessibility of data*: And this is a little person is leaving the organisation but they're taking their data with them and not putting it in the data centre to share with other people. And this person has had enough. They can't get the data they want. They're really **angry**.

*Insider threats***:** This, to me, I think was about recognising that there's an insider **threat**, both a witting insider **threat**, malevolent, but also an unwitting one.

*External threats*: So each of the people, people who are attacking the bits of technology that we've got. So we've got somebody who's halfway up the cloud, and he, but he's not at the top because it's secure. So he's **attacking** it, but he can't get there.

*Security breaches*: So I think it's just about being constantly aware of potential phishing and potential breaches and things like that. So it's like a work in progress. But I don't think, I don't personally think that there seems to be any one big **threat**.

*Cloud data*: I think that basically, the cloud is under the most **attack**. So there's the probably the most competent people are, if they want to try and get something, they're trying to **attack** the cloud. But it's also protected by the most competent people, in my opinion.

The co-text surrounding the word *angry* most often pertains to data issues whereas *threat* is associated more with the internal and external risks. The above extracts from the DDRC workshop transcripts are representative of spoken language and may be grammatically incorrect but they are illustrative of the diversity of the stories told by participants. They also display the openness that workshop attendees spoke with when sharing their attitudes towards data during the discussions.

## DISCONTENT

Similar to the sentiments expressed in the *Angry* domain, examining the co-text in the emotive *Discontent* category reveals the issues participants consider to be a hinderance to personal and organisational success. The only word that occurred more than once in this category is *frustrate*; inflected forms include *frustration* and *frustrating*. By examining the stories told by participants, a common narrative emerges which highlights perceived barriers to success. As shown in Figure 3, the sector that spoke the most about issues pertaining to discontent is the Government Department.

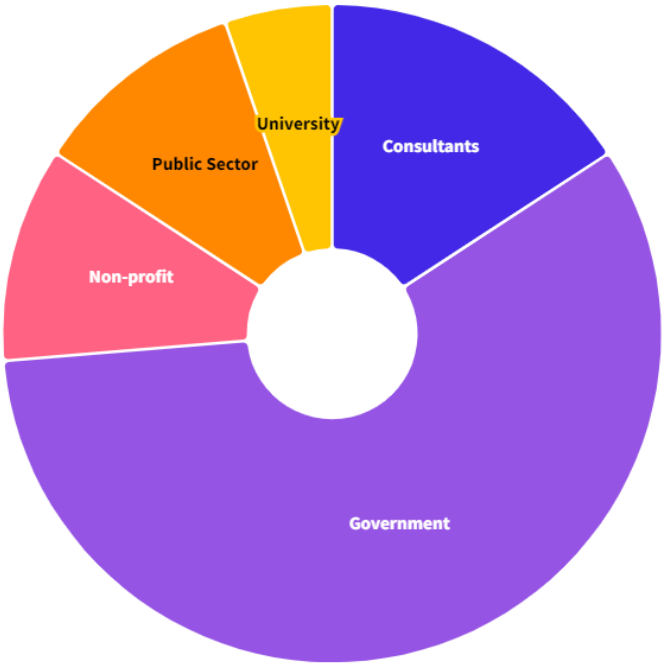



**Figure 3: Discontent Theme by Sector**

As shown, the Enterprise sector did not use language that conveyed dissatisfaction or obstacles in achieving success. The expressions of discontent within the Government Department primarily focused on challenges of data documentation, accessibility, storage, and transfer. Below are extracts from the DDRC workshop transcripts illustrating various stories told by participants of their concerns for barriers to success.

> *Duplication of effort*: It's **frustrating**, but it is our jobs. I guess. I'm also going to add to the part you said about I think a lot of searches happen so that we don't replicate work. So we do searches when we get to do things to do to make sure no one else has done it before or to find out information from them to then learn from and build on. And if those searches aren't clever enough, there's then risk that we waste money and do things again.

> *SharePoint accessibility*: This is Jake, who has got a red light on his computer, and he's not getting the file because it's being blocked by SharePoint Accessibility, **frustrations**, hence the wall between us.

> *VPN downtime and lack of support*: …that could happen to anyone working from home, like their access to Yeah, for a VPN might not work, or just general technical issues. And then that can only be **frustrating**. it's **frustrating** because there's no immediate point of contact to fix it, or it takes time to fix or at the time, and it is just an inconvenience, like you just want to get on your laptop and just work.

*Accessibility of data and availability of appropriate applications*: So then the rest of the pieces sort of about other problems that are sort of similar with respect to us transferring data within the organisation. So for example, like not being able to write papers jointly online, but having to use specific tools that aren't particularly well designed for that task, and the **frustration** that goes with that. So it means kind of like physically, you end up transferring data a lot of the time rather than being able to use cloud services and stuff like that because of the security associated with it.

*Accessibility of data:* So it teases you know that the data is there, but you know, you can't get it, which is the **frustration** so you're kind of like, you know, losing your head, like losing your cool over there.

*Replicating work due to permissions*: I was gonna say expectations are quite low sometimes. So it's **frustrating** but not always that it's not surprising. And it does depend, to your credit, it does depend on what type of data we're talking about here as well. So if I think as a social scientist, it is, I am more aware of what social science data we have, but less able to access it due to permissions.

The University sector expressed the least frustration but did mention issues with VPNs and inaccurate documentation as illustrated by the extract "Well, yeah, I mean, that was a specific example of we were frustrated with different versions of fact tables or dimensions or whatever. And we never really know which one we should be using." The concerns expressed in the domain of *Discontent* by DDRC workshop participants highlight a range of difficulties experienced in achieving success, but most comments focus on the accessibility of data.

## WORRY

The four most frequent root words that occur in the stories expressed by DDRC workshop participants for the semantic domain of *Worry* are *concern*, *worry*, *insecure*, and *care*. The distribution amongst sectors for all words identified as belonging to the theme of *Worry* is shown in Figure 4.


**Figure 4: Worry Theme by Sector**

Of the three semantic domains, *Worry* is the most evenly distributed across the six sectors, with non-profit containing the highest percentage. Collocates are words that occur immediately to the left or right of a word (typically 5 positions). The words *data*, *privacy*, *confidentiality*, *integrity*, and *priority* are five of the most frequently occurring collocates for the root words *concern*, *worry*, *insecure*, and *care*. Below are extracts from the DDRC workshops illustrating some of these issues:

> *Data integrity*: And the integrity of that data we felt was, I guess, a little bit of a ***concern*** because it comes from different members of the organisation, and it can feel a little bit chaotic.
>
> *Privacy of data*: Sometimes the privacy ***concerns*** caused a lot of Yeah, like, difficulties along transmitting data to each other.
>
> *Hybrid working and wellbeing*: They are really **worried** that something falls through the cracks for the system. I don't know, just from a wellbeing perspective, if we were all in the office together, then we would see which colleagues were really struggling with actually **worrying** about that.
>
> *Hybrid working and accessibility to data*: So now I am dependent on other people in the team, who are all very good and the software, to come up with the goods. And the **worry** for me is if the challenge me is if they're not around when I need that.

*Target marketing*: You know, the fact that you've mentioned something and all of a sudden there's a, there's an advert on your socials for that. So that **worries** me as an individual.

*Security and privacy of data*: Some data do set on cloud but not everything sits in cloud. There is a **concern** about privacy.

*Cyber-attacks:* Cyber-attacks are happening every day, on all the servers so at least everybody has to figure it out and not allow any kind of an error like impersonation attacks as well. So somebody tries to come in. So all those attacks, we have to be very **careful**.

*Security of legacy data*: And then this last one is more of a risk than the around raising awareness, but it's the idea of data sitting and rotting. So it's data that's sitting there, it's not secure. People are not **concerned** about it because may be mainly because they've forgotten about it.

*Security of data*: You're not **concerned** about the data because you think it's secure. Regulated. But I also think it's very high priority too. I mean, my financial data is very high priority.

The discussions of workshop participants that are categorised in the domain of *Worry* varies more than the other two emotive categories. There are employee issues such as hybrid working and wellbeing; security and technology risks like cyber-attacks; and numerous concerns were expressed regarding the privacy, integrity, and security of data.

## EMOTIVE CONCERNS SURROUNDING DATA

By combining the stories expressed by DDRC workshop participants in the co-text of the three marked emotive domains, a picture emerges of overarching concerns surrounding data. In the category of *Angry*, participants expressed not only the emotion of anger but also their perceived risks. Barriers or challenges to personal and organisation success are illustrated in the extracts for the domain of *Discontent*. Most of the discussions surrounding the domain of *Worry* focus on concerns relating to data, privacy, integrity, and security. Through a detailed examination of the text surrounding the high frequency words in these overarching themes, four major categories are revealed: data, technology, organisational processes, and employee. The target of these emotive concerns expressed by participants for each of the four categories are presented in Figures 5 – 7 and Table 3.

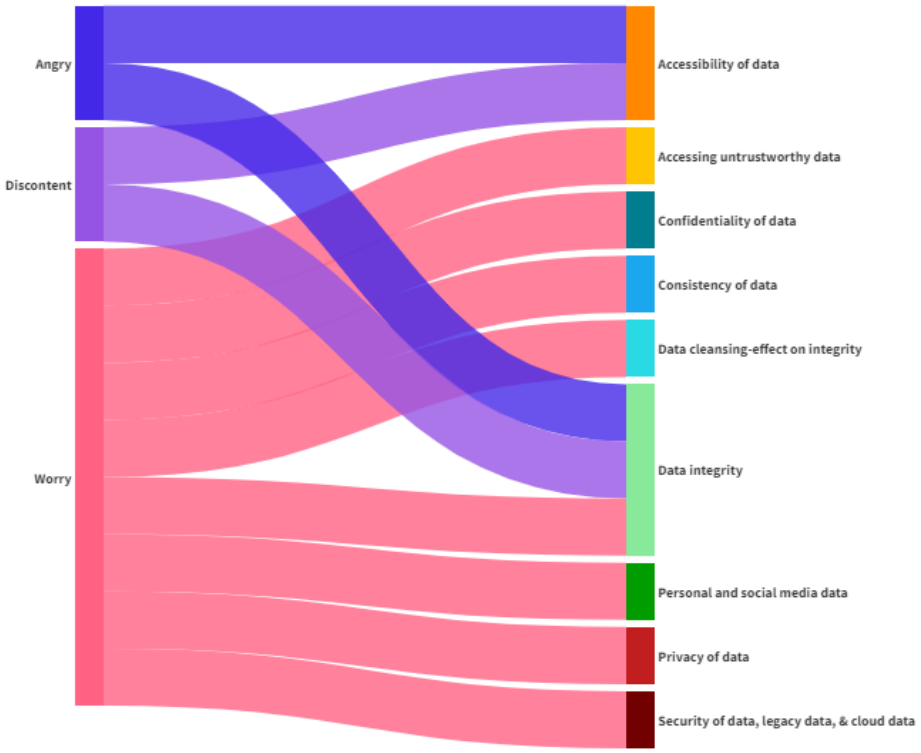



**Figure 5: Data Emotive Concerns**

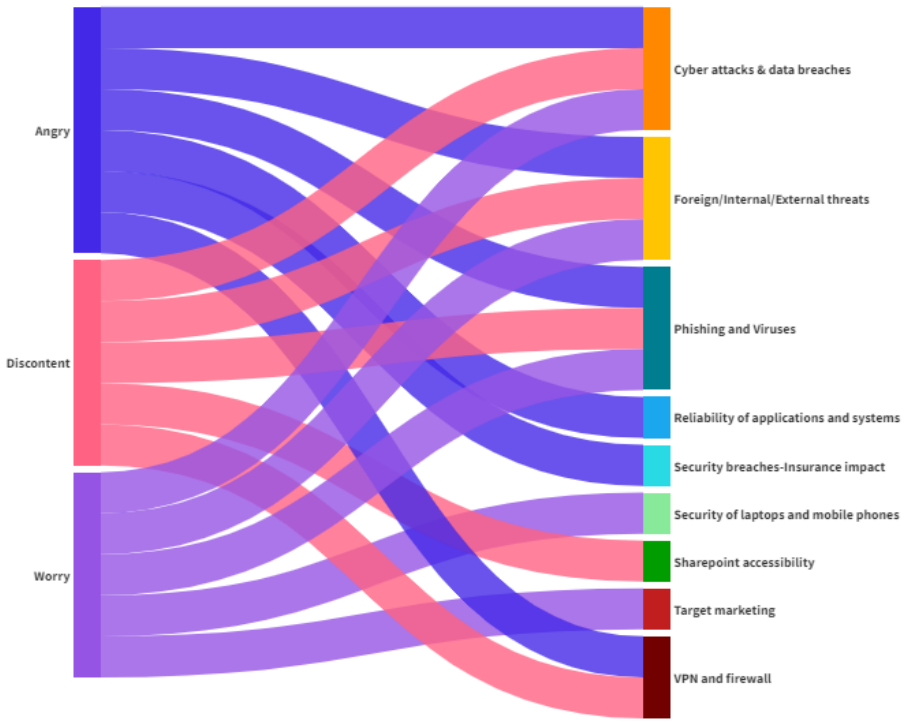



**Figure 6: Technology Emotive Concerns**

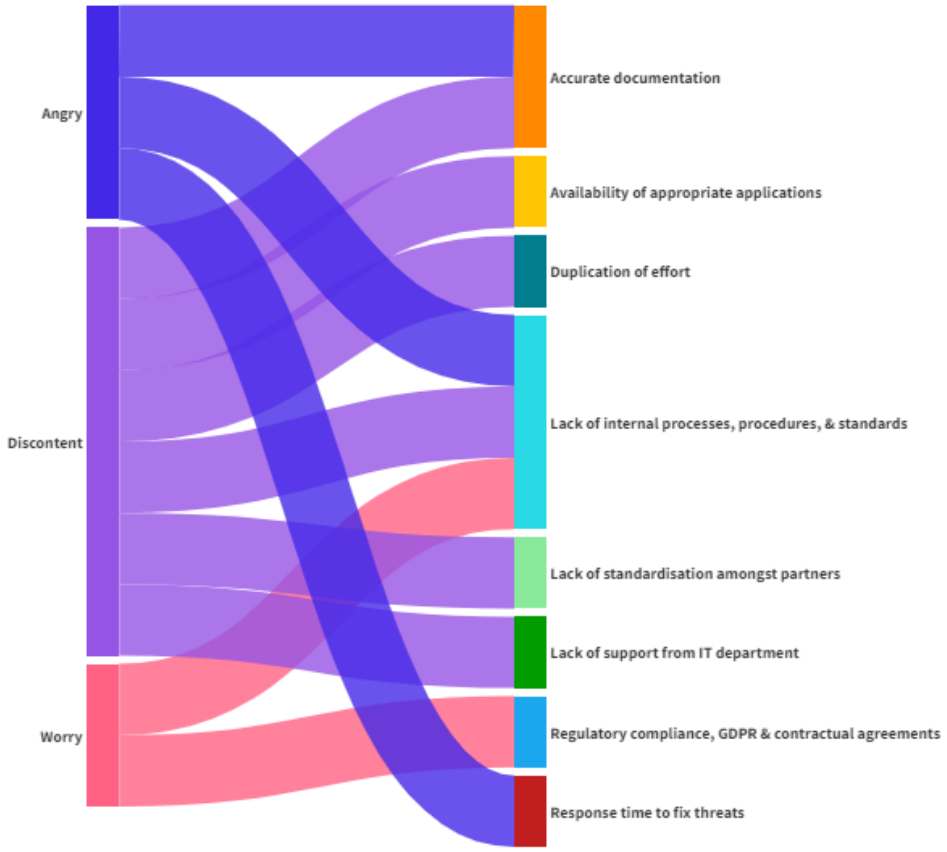



**Figure 7: Organisational processes emotive concerns**

| Emotion | Concern |
|---|---|
| Angry | Employee incompetence |
| Worry | Hybrid working-data access & wellbeing |
| Worry | Training on applications |
| Worry | Workload and the effect on data integrity |

**Table 3: Employee Emotive Concerns**

Figure 5 shows that data integrity is associated with all three emotive categories and accessibility of data concerns are related to discussions in the *Angry* and *Discontent* domains. The most common emotive data category is *Worry* which is mapped to everything except accessibility, implying perceived data risks are the main concern. Figure 6 illustrates that the three emotive categories as they pertain to technology are

nearly evenly distributed with *Angry* being slightly more common. There are three concerns that are linked to all three emotive domains, highlighting their importance to DDRC workshop participants: cyber attacks and data breaches; threats from foreign, internal, and external forces; and phishing attacks and viruses. As shown in Figure 7, the domain of *Discontent* is associated with every organisational concern except two, implying that barriers to organisational and personal success were an overarching theme in the DDRC workshops. The emotive concerns for employees are focused on hybrid working, training and employee workloads and the possible negative effect on data integrity.

It can be stated with certainty that the emotive concerns and issues presented above are representative of the stories told by participants attending DDRC workshops. There are shared and overlapping preoccupations and these issues should not be considered sector- or organisation-specific. The concerns expressed through emotive words such as *angry*, *frustration*, and *concern* pertain to not only personal feelings and emotions but also perceived risks and threats, and barriers to personal and organisational success.

## CONCLUDING REMARKS

The computational linguistic approach used in this report has provided a means of interpreting the stories being told by DDRC workshop participants and identifying marked themes. Quantitative analytical software reveals the three overarching emotive concerns in the transcripts are: *Angry*, *Discontent*, and *Worry*. These issues are examined at the word and co-text level for high frequency words within each category. An analysis reveals that the repeated language patterns can be categorised into four specific areas of focus: data, technology, organisational processes, and employee. A foundation is provided in this report for exploring the relationship between participants' language and the broader patterns that highlight their emotions surrounding data.

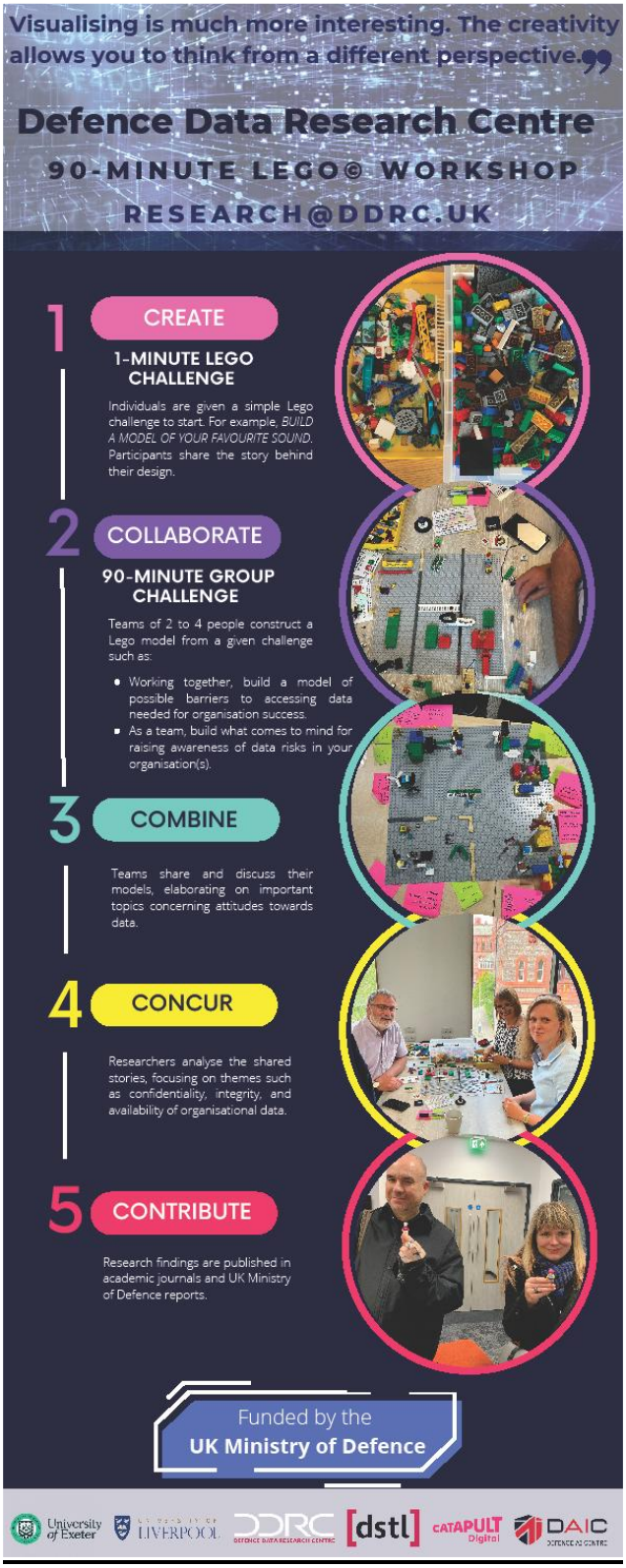
## REFERENCES

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.. The Sketch Engine: ten years on. Lexicography, 1: 7-36, 2014. Available at: http://www.sketchengine.eu.

Rayson, P. (2009) *Wmatrix: A web-based corpus processing environment*. Lancaster: Computing Department, Lancaster University. Available at: http://ucrel.lancs.ac.uk/wmatrix/.

Rayson, P. (2019) *Wmatrix for forensic linguistics: a practical hands-on demo*. Lancaster, England: Lancaster University. Available at: http://ucrel.lancs.ac.uk/paul/

## APPENDIX 1: RECRUITMENT INFOGRAPHIC



## APPENDIX 2: WORKSHOP TRANSCRIPTS WORD COUNT

| SECTOR | Edited Transcript | Original Transcript |
|---|---|---|
| **Consultants** | **9,254** | **9,484** |
| Another Day Consulting, London | | 2,418 |
| Frazer Nash, London | | 7,066 |
| **Government Department** | **30,927** | **31,348** |
| **Enterprise** | **16,982** | **17,038** |
| Naimuri, Manchester | | 3,317 |
| SME and OAP, Various Locations England and Wales | | 13,721 |
| **Non-profit** | **16,823** | **17,342** |
| CWMPAS, Cardiff | | 9,868 |
| Digital Catapult, London | | 4,911 |
| Turing Institute | | 2,563 |
| **Public Sector** | **23,397** | **27,401** |
| Manchester City Council Civil Service | | 12,166 |
| Newcastle Civil Service | | 8,184 |
| Scottish Government | | 7,051 |
| **Universities** | **27,696** | **28,175** |
| Turing Institite and University of Exeter Data Study Groups | | 8,613 |
| University of Exeter Professional Services | | 15,132 |
| University of Liverpool Professional Services | | 4,430 |