DDRC WORKSHOPS: IDEALISATION OF DATA SURROUNDINGS

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. A brief overview of the workshops carried out by DDRC is provided along with statistical confirmation that the project objectives were successfully met. The focus of this report is on the identification of common themes and high frequency words within these categories, as expressed in the language of those attending the workshops. Research findings on two significant themes are presented along with transcript extracts. The report findings identify similar language patterns across sectors, resulting in a vision for the ideal surroundings of data. This is primarily a quantitative report offering significant insights of prominent themes based on the language patterns of DDRC workshop participants.

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 (singular form of data), 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.

The DDRC facilitated 33 workshops between June and November 2023. A listing of participating organisations by sector and 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.

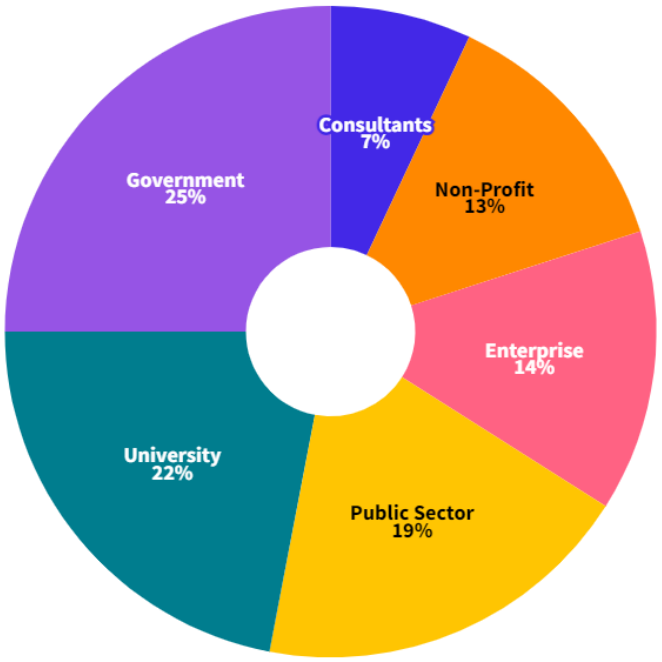


Figure 1: Sector Composition Based on Word Count

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 categories for a word's thematic concept (also known

as semantic domain). In this analysis, only the first semantic domain identified by Wmatrix is used, offering a 91% accuracy rate for a given word in context (Rayson 2019).

PROMINENT THEMES AND NEGATION

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 two of these: *Obligation or Necessity* and *Positive Evaluation*. High frequency words within both domains can aid in the identification of an idealised vision for data surroundings. The two key thematic concepts show a near equal distribution by sector, representing similar lexical and thematic choices made by the workshop participants. In what follows, 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. A breakdown of the words employed by participants in each sector for the two marked categories are shown in Figures 2 and 3.

Figure 2: Obligation or Necessity

Figure 3: Positive Evaluation

The graphs show nearly equal distributions across sectors for both domains. 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, signaling an array of concerns surrounding data. The marked words in descending order of frequency in the *Obligation or Necessity* domain include but are not limited to: *need*, *should*, *necessary*, *essential*, and *must*. The key words for *Positive Evaluation* include in descending order: *trustworthy*, *good*, *great*, *better*, *brilliant*, *perfect*, and *ideal*. 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' vision for successful policies, practices, and procedures as they relate to data.

It is important to note that often the high frequency words in these two categories can be negated. Examples from the workshops include:

> You think he's trustworthy, it's trustworthy, **but actually** here he is leaking the data out here, this red, this red pipe represents the leakage of data out to external actors.
>
> You know, and what we discussed was is, no matter how good the data is, if you've got the risk of having low priority information, and certainly **untrustworthy** data coming through, you completely lose confidence in what you receive.

In the first example, *but actually* negates the word *trustworthy* but the meaning is significant because an ideal of not leaking data to external actors is presented. Similarly, in the next extract the prefix *un-* transforms *trustworthy*, indicating issues that can lead to a lack of confidence in data. Lexical issues such as negation are accounted for in the research findings.

OBLIGATION OR NECESSITY

By examining the transcript research dataset for the phrases and sentences surrounding high frequency words in the semantic domain *Obligation or Necessity*, a narrative emerges of DDRC workshop participants' preoccupations with what is needed, necessary and essential. Patterns are identified from the stories told and four distinct categories of concern emerge: organisational, technology, data and employee. These are shown in Figure 4 along with the individual topics that were discussed across the six sectors.

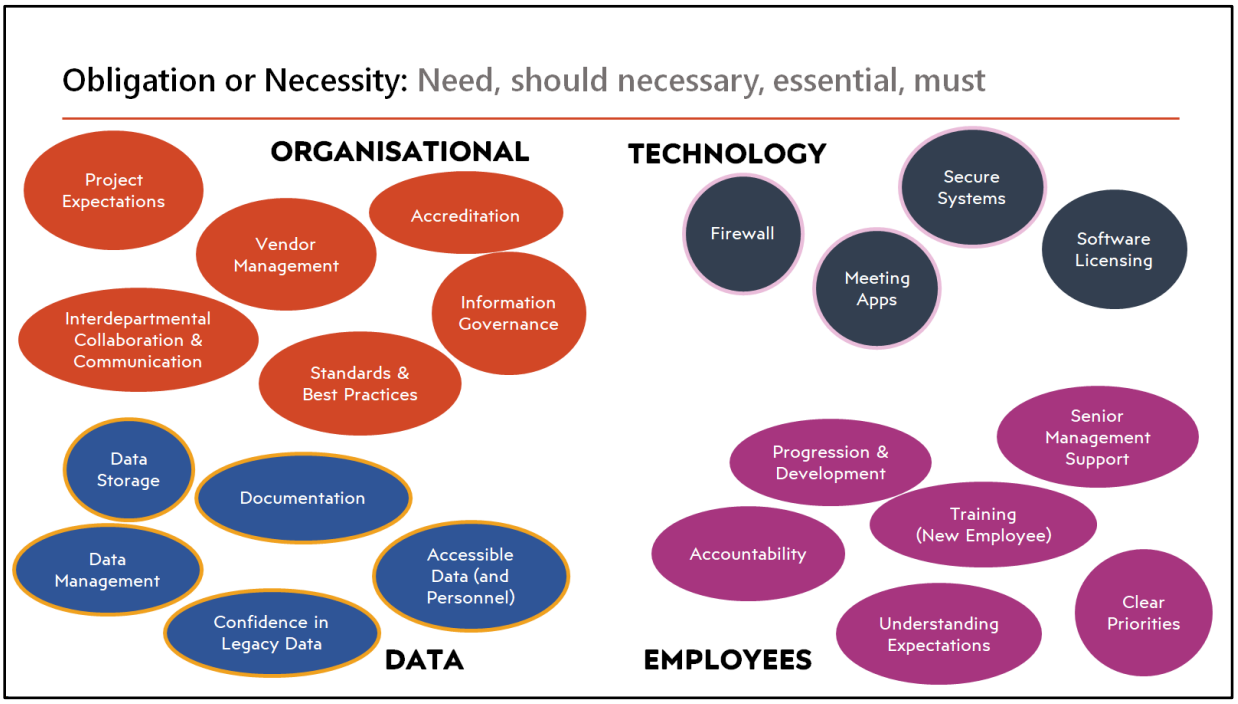


Figure 4: Obligation or Necessity Concerns

The concerns expressed in Figure 4 were shared 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).

> *Vendor Management*: And so, the eyes there represents some kind of us monstrosity examining all the relationships we have with vendors. Are they really **necessary**? Are they up to date? Do they still serve the function they were intended to? Do they still serve the function they were intended to? Just putting them under a microscope really to see if we **should** continue or not.
>
> *Understanding Priorities and Expectations*: Because we've got all these requirements, we can't deliver everything. We have to make sure that we understand what are the things we **must** deliver. Why is it important? How is it a priority?
>
> *Accessibility to Data*: There's a reason why it **should** be difficult to access for the wrong people, but we're making it difficult to access for the right people.
>
> *Documentation*: 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.
>
> *Interdepartmental Communication & Data Documentation*: You know, setting out clear definitions of what the data **should** be, and how it's stored, who's owning that data. And so that can be communicated down.
>
> *Standards and Best Practices & Data Management*: I think it's about all approaches to general data housekeeping. And there's no one model of - this is how we **should** do it. And there's no championing of best practices. And that's something that we all want. So that we know what the standards **should** be.
>
> *Information Governance, Interdepartmental Communication & Data Storage*: We do have, like a Research Services team, which means when or when a grant comes through, they **should** be looking over the data they're going to be using is secure. We **need** them to because I'm the IT team, I support the system that they're using. It's not my job to **necessarily** educate them. This is what you **should** be doing with data. That's the research services or information governance team **should** be doing that. Obviously, I go to somebody's desk to help them with a problem and I see they're working on this data which **shouldn't** really be on their own laptop, you know, on a USB stick on their desk or something like that, you know, **should** be stored securely.

The use of key words such as *need* and *should* illustrate important and desirable facets such as documentation and communication as they pertain to data The above extracts from the DDRC workshop transcripts are representative of spoken language and may not be easy to read but they are illustrative of the diversity of the language employed. They also display the openness that participants spoke with when sharing their attitudes towards data during workshop discussions.

POSITIVE EVALUATION

Similar to the above analysis, examining the co-text of high frequency words in the *Positive Evaluation* semantic domain reveals what participants consider trustworthy, good, and ideal in organisational settings surrounding data. A common narrative emerges, highlighting three critical areas of concern: technology, data, and organisational. As shown in Figure 5, the most often discussed theme is the identification of ideal organisational practices and procedures as they pertain to data.

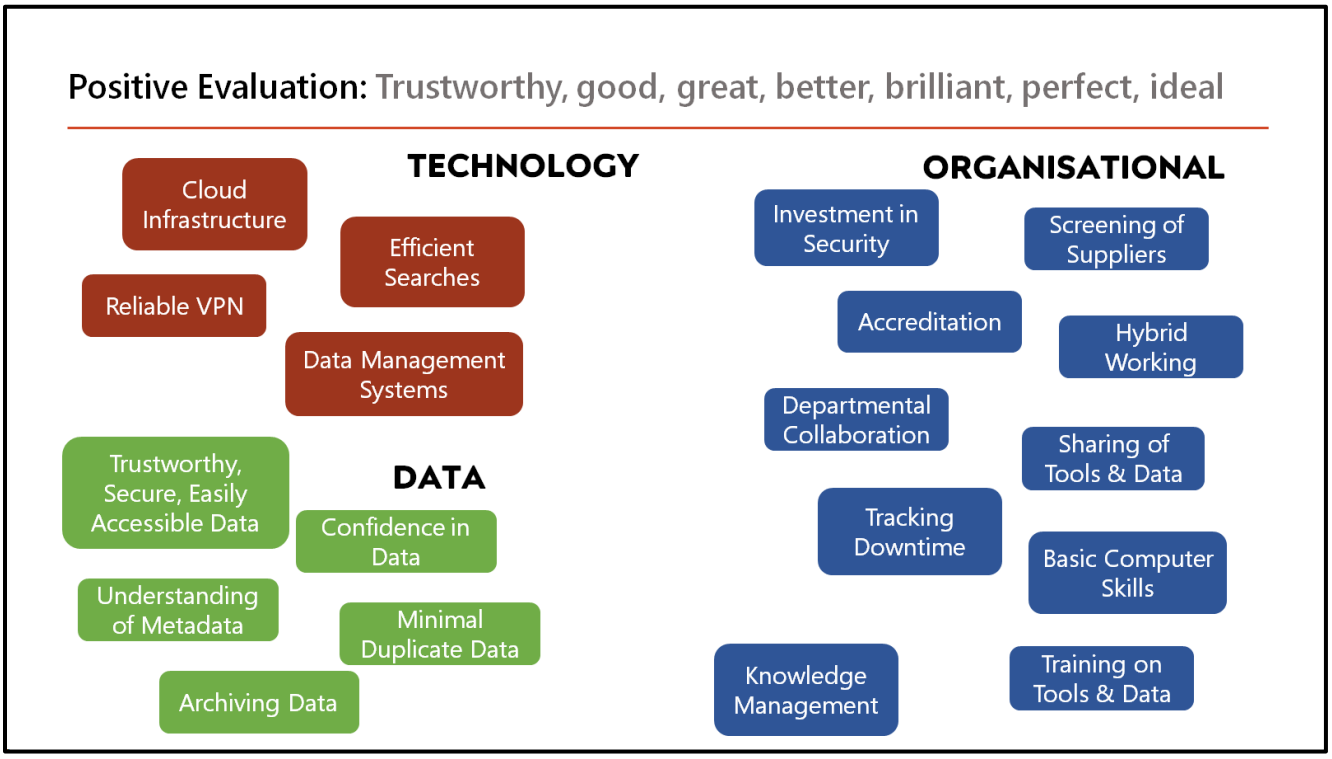


Figure 5: Positive Evaluation Concerns

Below are extracts from the DDRC workshop transcripts illustrating various stories told by participants of idealised processes and procedures pertaining to technology, data, and organisations.

> *Secure Data & Investment in Security*: Because he wants to keep the business reliable and **trustworthy**. Or if you want to cut a few corners, you can take the money and make sure that there's a lot of data breaches.
>
> *Cloud Infrastructure:* And we've got **good** kind of internal and external cloud infrastructure, some of which I think, I believe is, you know, very, very **trustworthy** and secure.
>
> *Departmental Collaboration & Trustworthy Data:* We cleaned up some of the not **trustworthy** and some of the low priority data, and some of the stuff that's less useful. And then at this point, it goes into separate directions to come back to us. So it goes to experts. And we do some processing as well. So it would if this was the **ideal** world thinking.
>
> *Secure Data & Trustworthy Data*: Yeah, so in the **ideal**, you'd have your **trustworthy**, and secure data, which is then easily accessible to those that need it.
>
> *Knowledge Management, Departmental Collaboration & Accessible Data*: So we felt, in terms of actually finding improvements, streamlining a lot of the bureaucracy, both for sharing and for accessing internally, and **better** knowledge management, would be a really good way to improve things.
>
> *Training*: The training, it's still training. But yeah, boy it's one of those things where, Oh, you can ask someone that you're close to, you have a **better** relationship with, and not feel embarrassed?
>
> *Confidence in Data*: You know, and what we discussed was is, no matter how **good** the data is, if you've got the risk of having low priority information, and certainly **untrustworthy** data coming through, you completely lose confidence in what you receive. So actually, you start questioning what you receive and all the **good** stuff start to get ignored.
>
> *Screening of Suppliers:* I think for example, when we look at supplier data, we go through a due diligence process for them to become a part of our supplier network. So the data that we get is **trustworthy**.

Words such as *trustworthy*, *good,* and *brilliant* were employed frequently by workshop participants in expressing individual concerns surrounding their personal experiences in their organisational settings. These extracts illustrate the commonality of workshop discussion topics across sectors as shown in Figure 3.

IDEALISATION OF DATA SURROUNDINGS

By combining the stories expressed by DDRC workshop participants in the co-text of the two marked semantic domains, a picture emerges of desired practices, processes, and procedures as they pertain to data. Through a detailed examination of the text surrounding the high frequency words in both overarching themes, four major categories are revealed: data, employee, organisational, and technology. Common concerns or ideals expressed by participants for each category is presented in Figure 6.

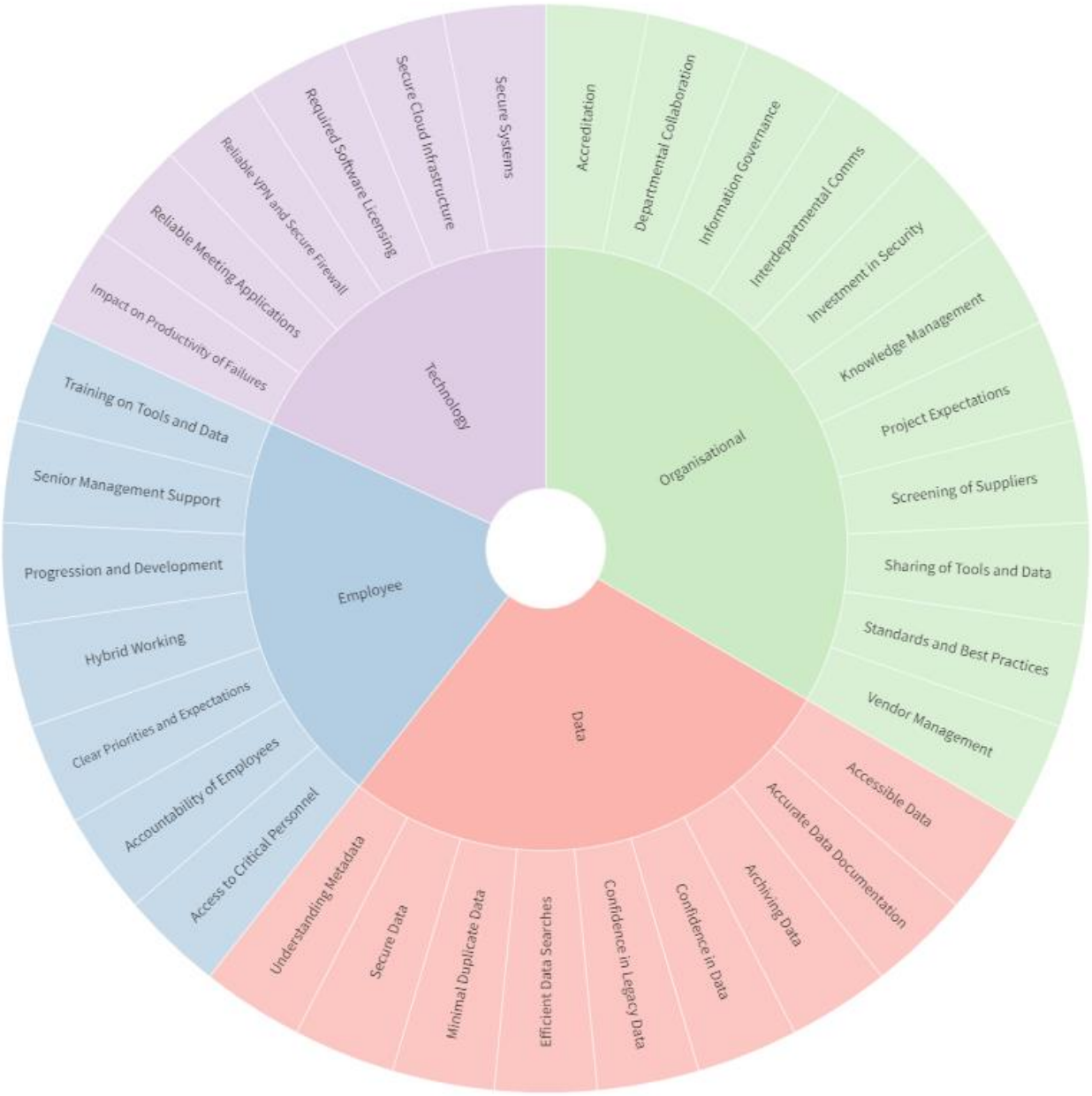


Figure 6: Idealisation of Data Surroundings

It can be stated with certainty that the issues presented above are common preoccupations of the DDRC workshop participants when discussing an idealisation of data surroundings. They should not be considered sector- or organisation-specific issues as comments across workshops were similar. These concerns can be viewed as an ideal of procedures and practices for organisational success as they pertain to data.

CONCLUDING REMARKS

The computational linguistic approach used in this report has provided a means of identifying and interpreting the stories being told by DDRC workshop participants. From an objective perspective, significant patterns in the research dataset are identified which may not be discernible through close reading alone. The language employed by participants is reported on and signals similarities across sectors for two overarching themes: *Obligation or Necessity* and *Positive Evaluation*. These are examined at the word and co-text level, revealing dominant concerns expressed in the DDRC workshops. A foundation is provided for exploring the relationship between participants' language and the broader patterns that highlight their preoccupations with the environment of data. An analysis of these discussions reveals four categories of common concerns across the six workshop sectors, resulting in a mapping of the ideal surroundings for organisational data.

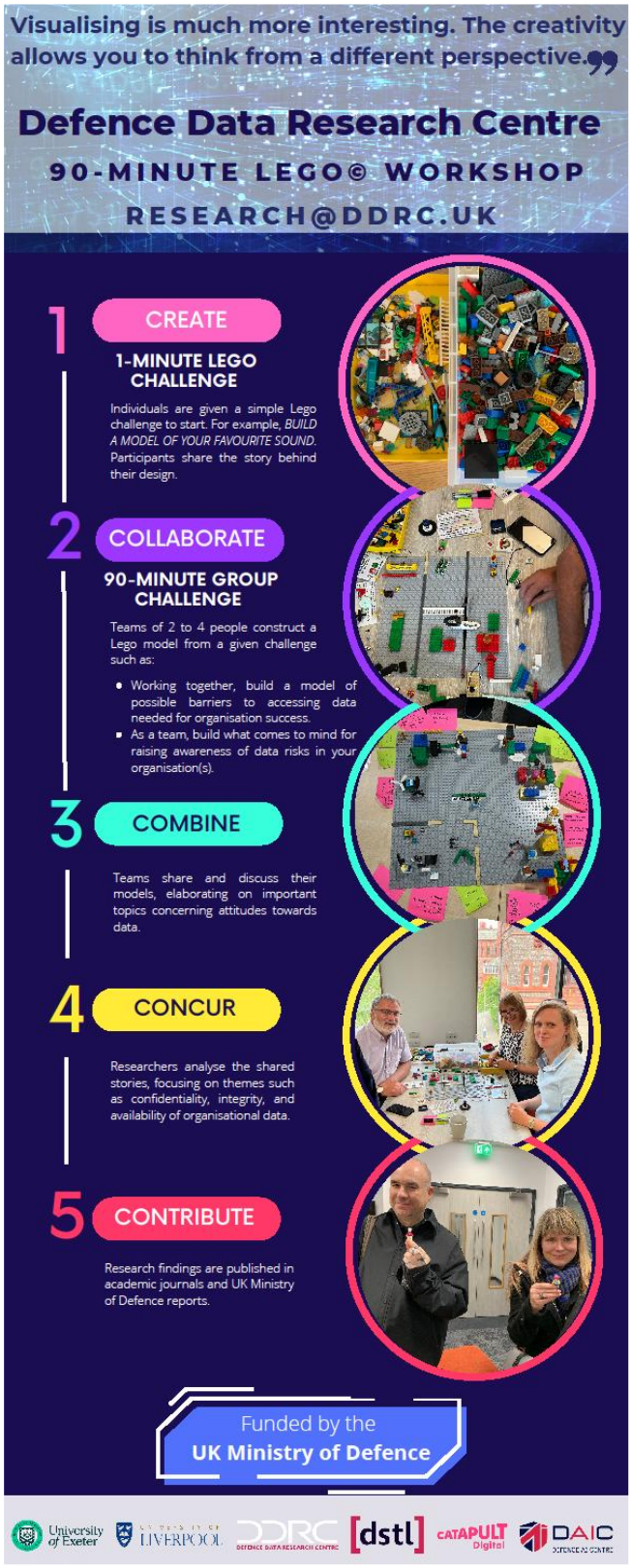
REFERENCES

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: WORKSHOPS BY SECTOR WITH 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