FDS Briefing note 6

Data Access Spectrum

Authors: Felix Ritchie and Cara Kendal, UWE DRAGoN

In some cases, presenting the five dimensions of data safety (people, projects, settings, data and outputs) as defined by the Five Safes Framework can be confusing and difficult to depict. The Data Access Spectrum is a popular way to depict alternative data access solutions.

The depiction hinges on the understanding that four of the five safes (people, projects, settings and outputs) typically work together when managing risk and can be effectively understood as a set of procedural controls. These procedural controls fit to data as the remaining dimension alongside technical setting, with the level of procedural control depending on the detail and required anonymity level of the data. This allows us to depict the data access spectrum as a linear scale rather than a multidimensional graph. Figure 1 presents a version commonly used in the UK.

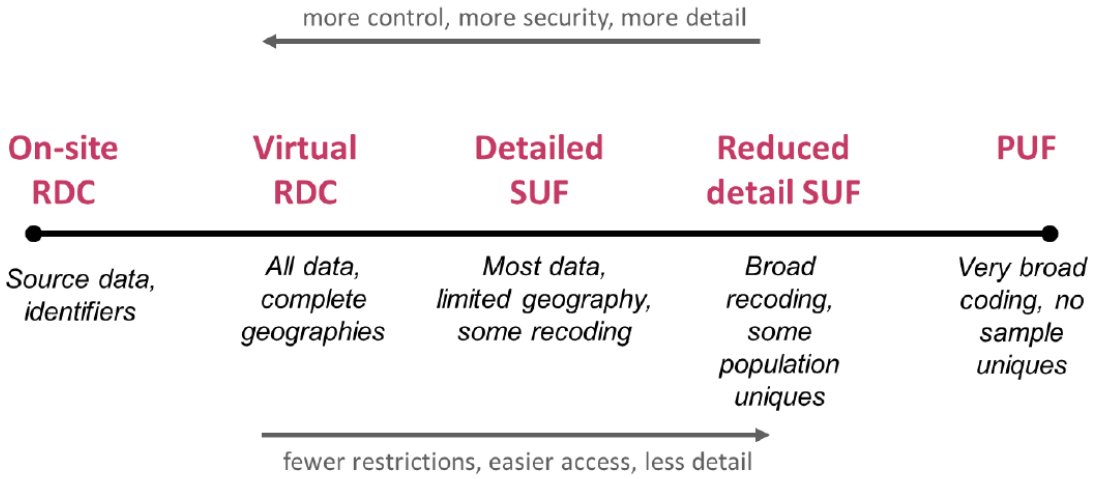

Figure 1 Data access spectrum from Green and Ritchie, 2016[1]

This scale indicates the increased need for procedural controls and security measures are data gets more detailed and therefore gains greater potential to be identifying. The way that we interact with and protect data along this spectrum varies.

Specific example - LFS

This model is called the 'continuum of access' in Canada where it was simultaneously and independently developed by Chuck Humphrey of the Canadian Research Data Centres Network. This model has also been promoted by other organisations such as the Open Data Institute (ODI), although whether as an innovation or as a development of the other models is not clear. It seems likely that such a simple but useful device would be re-invented multiple times.

[1] ADSS paper

The ODI representation is slightly different, and is a 'data spectrum' rather than 'access':

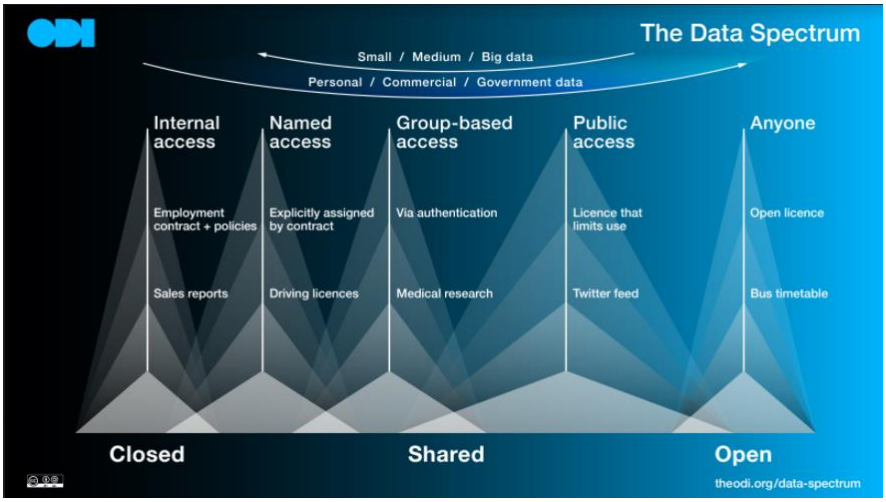

Figure 2 The ODI Data Spectrum (source: www.theodi.org)

This may reflect the ODI's stronger focus on private sector organisations, where contractual agreements and commercial considerations play a larger part than in public sector data models.

Differences along the Data Access spectrum

We can use the DAS in multiple ways to examine differences between data solutions. We consider six options:

- Source data: primary data collection, which may include direct identifiers
- Secure Use: access to researchers is only through a facility controlled by the data holder
- Certified download: researchers can download data subject to a formal application and review by a human
- Self-certified download: researchers can download data subject to automatic checks (eg academic email address) and making commitments to standard terms (eg not copying to portable media, deleting at project end)
- Open data: no restrictions on downloading and use
- Aggregate data: data represents statistical characteristics of respondents

Aggregate and Microdata

One way of distinguishing between stages of the data access spectrum is to consider where aggregate data and microdata fall along it.

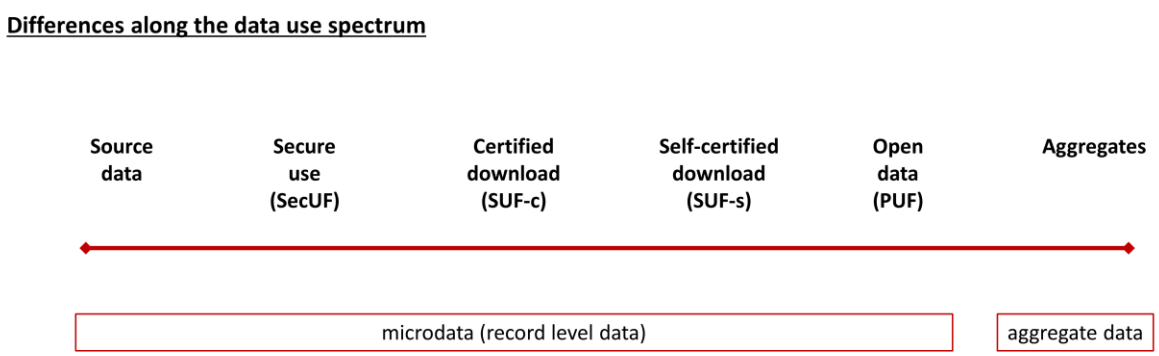

Figure 3 Aggregate data and microdata on the data access spectrum

As the above depiction of the data access spectrum demonstrates, aggregate data and microdata/record-level data differ significantly in their detail, sensitivity, and the procedural controls required for their management and access. Aggregate data is a summarised form of data that combines individual records to produce statistics or totals, reflecting broader trends or patterns within a dataset. Examples can include averages for a region or total observations within a timeframe. In contrast, microdata, or record-level data, consists of detailed records containing information about individual 'records' such as people, organisations or households. Therefore, we position aggregate data towards the very lowest end of the data access spectrum as it is less sensitive and less detailed allowing it to be generally more accessible and of lower risk of revealing personal information. Due to its generalized nature, aggregate data can often be released publicly or shared with minimal restrictions, providing general insights without compromising privacy. On the other hand, microdata is more detailed and sensitive, requiring more rigorous controls.

Distributed Data vs Distributed Access

Another useful way to consider the data use spectrum is the line between distributed data and distributed access.

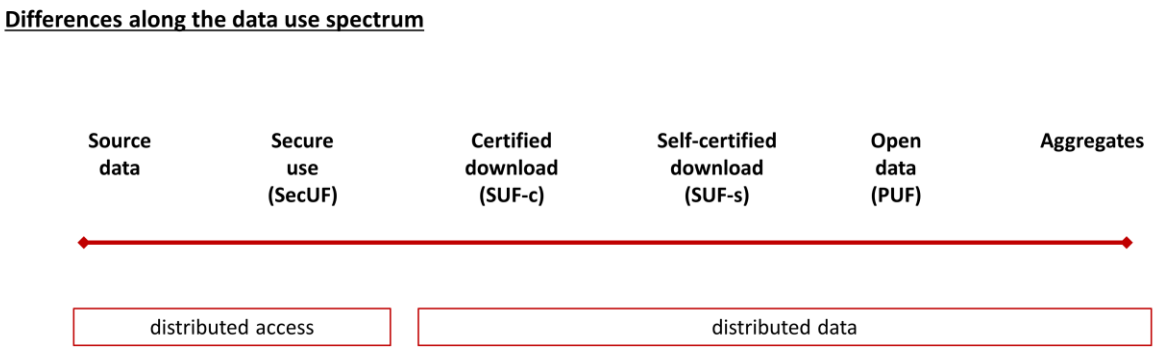

Figure 4 Distributed data and distributed access on the data spectrum

Distributed data and distributed access to data represent different approaches to data dissemination and control along the data access spectrum. Distributed data refers to datasets that are made widely available, often in a less detailed form, for public or semi-public use. These include public use files, self-certified downloads, and certified downloads, which are typically anonymized or aggregated to minimize privacy risks and ensure broader accessibility. The ease of access and reduced sensitivity make distributed data suitable for broad dissemination. On the other hand, distributed access to data involves providing controlled access to more detailed and sensitive datasets, such as secure use files and source data. Access to such data is tightly regulated, often requiring researchers to view the data from a secure environment.

Identified, Deidentified and Anonymised data

The data types across the access spectrum can be understood also as a scale between identified, deidentified and anonymised data.

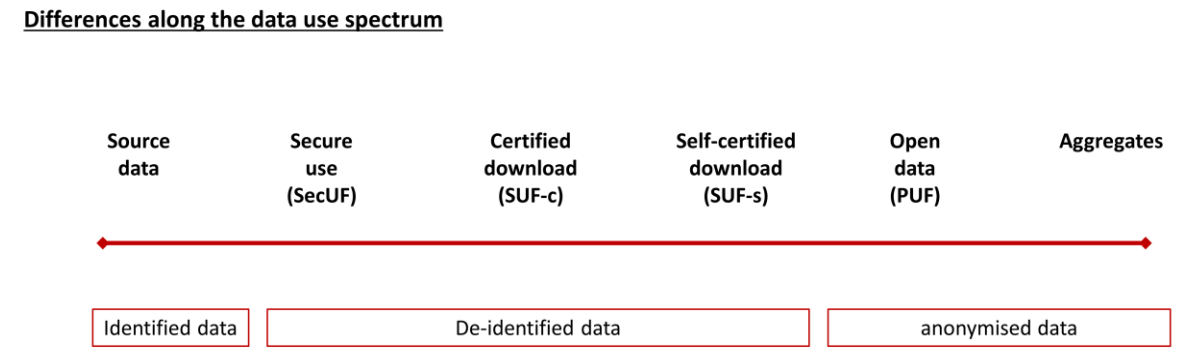

Figure 5 Identified, De-identified and anonymised data on the data access spectrum

Identified data, deidentified data, and anonymized data represent different levels of detail and privacy control along the data access spectrum, each reflecting varying degrees of sensitivity and accessibility. Identified data contains personal identifiers, such as names, addresses, or ID numbers, making it highly detailed and sensitive. They can also contain data that are not direct identifiers but are still likely to identify a subject due to subject uniqueness. This type of data, found in some secure use files and source data, provide detailed insights but therefore require more strict controls. In contrast, deidentified data has had identifiers removed, reducing the risk associated with the data, but may still include indirect identifiers that create a need for a degree of access regulation and procedural control. This data type includes self-certified downloads, certified downloads, and some secure use files. At the least detailed end of the spectrum is anonymised data which refers to data where all identifiers (direct and indirect) have been removed and there is no reasonable risk of an individual being identified from that data. This data is usually aggregates and public use files, providing the least detail and is generally the least sensitive. Therefore, this data type can be broadly distributed without significant privacy concerns.

The Five Safes along the Data Access Spectrum

The Five Safes framework can be brought back in to show how control is adapted across the DAS:

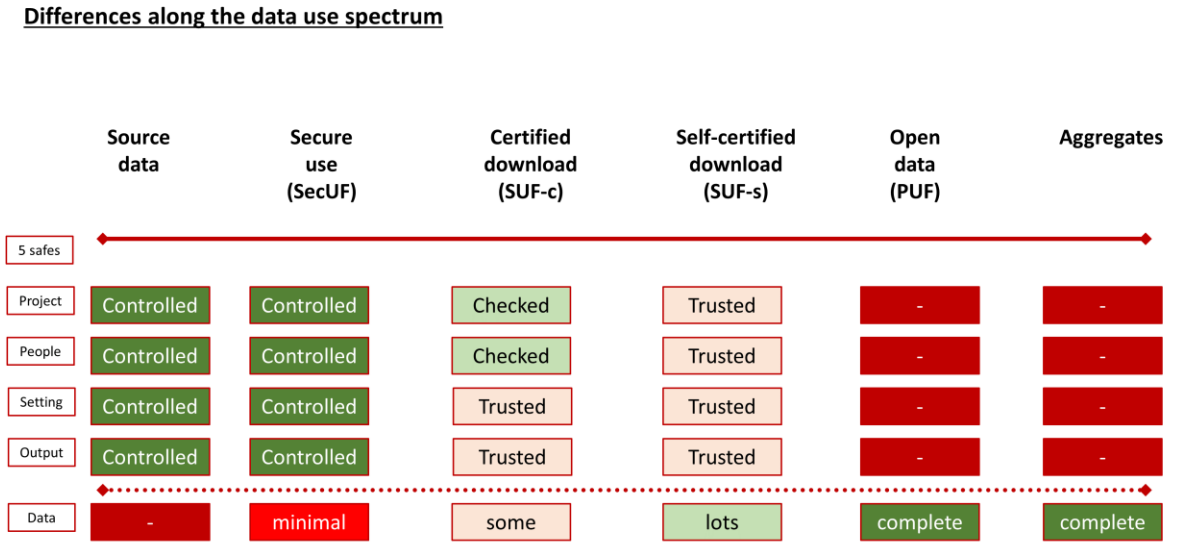

Figure 6 Five Safes controls applied to the data access spectrum

The Five Safes framework—comprising safe data, safe projects, safe people, safe settings, and safe outputs—dynamically adjusts along the data access spectrum to ensure appropriate data protection. At the least detailed end, with aggregates and public use files, the data is completely anonymized, making it inherently safe and thus negating the need for additional safeguards related to projects, people, settings, or outputs. Moving along the spectrum to self-certified downloads, the data remains mostly anonymized but may include some indirect identifiers, necessitating trust in the responsible individuals, projects, settings, and the eventual outputs to ensure continued protection. With certified downloads, where the data is somewhat more detailed, safety measures should incorporate trusted settings and trust in outputs, alongside checks regarding the projects using the data and the researchers and other people involved who will have data access. At the most detailed end of the spectrum, encompassing secure use files and source data, the data itself offers minimal inherent safety. Thus, comprehensive controls across all four other safes are important.