# The Colour of Offline Data

Brian Ballsun-Stanton ⬤*

Faculty of Arts, Macquarie University

Fieldwork results in the creation of data while offline in remote environments. This data may suffer from issues ranging from poor recording forms, a hostile recording environment for the collectors, or measurement uncertainty due to the sample condition. Data from fieldwork and any structured metadata cannot practically encompass all the in-field decisions, methodological choices, environmental factors, and compromises encountered by the field researchers. All these factors, however, still influence quality and future uses of collected data. Skala describes this context around the data as the colour of its bits: a human-evaluable context which operates much like a document's security clearance which is an emergent property of the context of the creation of the data (2004).

While some aspects of the context of the creation and history of the data should be captured in metadata during creation, updates, and deletion—the data model must anticipate these context-events and provide them their own data structure. Well-designed data structures do an excellent job of capturing salient and observable metadata events, especially if the data originates from within an online system. However, it is impossible to provide a full context and history of any given real-world observation represented as data. Skala uses the context of a recording of John Cage's 4'33":

> One of my friends was talking about how he'd performed John Cage's famous silent musical composition 4'33" for MP3. Okay, we said, ... so you took an appropriate-sized file of zeroes out of /dev/zero and compressed that with an MP3 compressor? No, no, he said. If I did that, it wouldn't really be 4'33" because to perform the composition, you have to make the silence in a certain way, according to the rules laid down by the composer. It's not just four minutes and thirty-three seconds of any old silence (2004).

---

*<brian.ballsun-stanton@mq.edu.au>

In my own experience, this colour *qua* unrecorded-context-around-data is most evident from data produced from archaeological fieldwork. To a 'colour-blind' computer scientist, where a bit is a bit, archaeological data collected in an app is either valid or invalid according to rules implemented in the recording form, program, or database structure. In the messy pragmatics of the field, there exist colours like 'not valid yet, look again after lab work', 'field supervisor should perform quality assurance first', 'physically uncertain,' and 'this model is not appropriate, but we have to record anyways,' amongst other operational nuances that cannot be well captured in metadata. Archaeological fieldwork tends towards a do-once activity, either because it is physically destroying the environment it is recording as it records the environment or due to the difficulties of obtaining data—high-quality data and record validity is a goal rather than a minimum threshold.

# 1 Metadata and the Universe of Discourse

The delineation of a boundary dividing model versus not-model is called the 'Universe of Discourse' by Venn: 'When we make use of names and resort to reasonings, what limits of reference, if any, do we make? What is the range of subject matter about which we consider ourselves to be speaking?' (Venn 1881, p. 180). Here, specifically, Venn describes the challenges of excluding all the irrelevant factors to our reasoning and describes the rectangular bounding box surrounding his famous circles as the Universe of Discourse – the division between 'these are matters we care about' and 'these are not matters of relevance to the current model.'

It is impossible to have a model with sufficient data and metadata to be a fully accurate representation of the thing without being the thing itself (see Borges 1975, p. 131 and the map-territory distinction). Therefore, the model designer must always anticipate a future universe of discourse when making data models and providing for metadata and annotation. There are always boundary trade-offs where the universe of discourse intersects the model. This constraint holds for a relational database as much as it does the trained eye and narrative articulated by Caraher in Slow Archaeology, 'Documenting features in a trench or in the field in handwritten notebooks provides a moment to slow down and to observe subtleties that we might have otherwise missed in our quest for efficient data collection' (2015, p. 50). Data/text description must elide non-salient detail outside the universe of discourse. A model or description which fails to do so decreases in utility as data entry takes more time for increasingly marginal gains.

# 2 Colours in Archaeological Fieldwork

Outside of Skala's colours informing the intellectual property status of a bit, or Patrick McKenzie's discussion around the accountant's view of various metadata-infused-colours of money (2022), there exist colours of academic data. In many ways, this mirrors the tacit knowledge carried by researchers and machines from lab to lab in a wordless pidgin—the data itself and metadata at the document or record level may be insufficient

to contextualise or reproduce the data without the original researchers or machines in a specific configuration (Galison 1997, p. 51).

Capturing this wordless pidgin, the relationship of the thing in context with its (complex) data-capturing environment, is outside the pragmatic realities of fieldwork. Those creating a record (digital or traditional) establish an implicit universe of discourse as they design their recording forms and methods with the intention that the records will serve to adequately capture the data under observation.

In this paper, I will use the implementation of 'annotation' and 'certainty' value-level metadata in an offline-focused field-data collection software app across more than seventy projects as a case study exploring how the metaphor of the colour of data illuminates the wordless pidgins of fieldwork, record preparation, and data analysis. Furthermore, I will discuss how data created offline, through remote fieldwork, is commonly 'stateful'— requiring distinct kinds and strengths of validation and metadata as it moves from field to lab to paper, even though the data itself may not change.

This paper should be able to assist in design discussions around software systems which may support offline or remote fieldwork. Some software developers may have the belief when designing systems that, '... on the other hand, bits are bits are bits and it is absolutely fundamental that two identical chunks of bits cannot be distinguished. Colour does not exist' (Skala 2004). I assert that they are wrong for these use-cases. Software systems which cannot handle the context and state of data, its 'colour,' as it passes through the research pipeline are ill-suited for remote or offline fieldwork.

## 3 Acknowledgements

## References

Borges, Jorge Luis (1975). *A Universal History of Infamy*. en. Penguin Books. ISBN: 9780140039597.

Caraher, William (2015). "Slow archaeology". In: *North Dakota Quarterly* 80.2, pp. 43–52. URL: https://hcommons.org/deposits/download/hc:11576/CONTENT/slow_archaeology.pdf.

Galison, Peter (Oct. 1997). *Image and Logic: A Material Culture of Microphysics*. en. University of Chicago Press. ISBN: 9780226279176.

Mckenzie, Patrick (Mar. 2022). *Accounting for SaaS and swords*. Accessed: 2023-3-1. URL: https://web.archive.org/web/https://www.bitsaboutmoney.com/archive/accounting-for-saas-and-swords/.

Skala, Matthew (June 2004). *What Colour are your bits?* en. Accessed: 2023-2-10. URL: https://web.archive.org/web/https://ansuz.sooke.bc.ca/entry/23.

Venn, John (1881). "The universe of discourse". In: *Symbolic logic*. Ed. by John Venn. Vol. 446. London, Great Britain: Macmillan and Co, xl, pp. 180–189. DOI: 10.1037/14127-009.