On the Nature of Information: How FAIR Digital Objects are Building-up Semantic Space

Hans-Günther Döbereiner [‡]

‡ Universität Bremen, Bremen, Germany

Corresponding author: Hans-Günther Döbereiner (hg.doebereiner@icloud.com)

Reviewable | v 1

Abstract

In this paper, we are concerned about the nature of information and how to gather and compose data with the help of so called FAIR digital objects (FDOs) in order to transform them to knowledge. FDOs are digital surrogates of real objects. The nature of information is intrinsically linked to the kind of questions one is asking. One might not ask a question or get philosophical about it. Answers depend on the data different disciplines gather about their objects of study. In Statistical Physics, classical Shannon entropy measures system order which in equilibrium just equals the heat exchanged with the environment. In cell biology, each protein carries certain functions which create specific information. Cognitive science describes how organisms perceive their environment via functional sensors and control behavior accordingly. Note that one can have function and control without meaning. In contrast, psychology is concerned with the assessment of our perceptions by assigning meaning and ensuing actions. Finally, philosophy builds logical constructs and formulates principles, in effect transforming facts into complex knowledge. All these statements make sense, but there is an even more concise way. Indeed, Luciano Floridi provides a precise and thorough classification of information in his central oeuvre *On the Philosophy of Information* (Floridi 2013). Especially, he performs a sequential construction to develop the attributes which data need to have in order to count as knowledge. Semantic information is necessarily well-formed, meaningful and truthful. Well-formed data becomes meaningful by action based-semantics of an autonomous-agent solving the symbol grounding problem (Taddeo and Floridi 2005) interacting with the environment. Knowledge is created then by being informed through relevant data accounted for. We notice that the notion of agency is

crucial for defining meaning. The apparent gap between Sciences and Humanities (Bawden and Robinson 2020) is created by the very existence of meaning. Further, meaning depends on interactions & connotations which are commensurate with the effective complexity of the environment of a particular agent resulting in an array of possible definitions.

In his classical paper *More is different* (Anderson 1972) discussed verbatim the hierarchical nature of science. Each level is made of and obeys the laws of its constituents from one level below with the higher-level exhibiting emergent properties like wetness of water assignable only to the whole system. As we rise through the hierarchies, there is a branch of science for each level of complexity; on each complexity level there are objects for which it is appropriate and fitting to build up vocabulary for the respective levels of description leading to formation of disciplinary languages. It is the central idea of causal emergence that on each level there is an optimal degree of coarse graining to define those objects in such a way that causality becomes maximal between them. This means there is emergence of informative higher scales in complex materials extending to biological systems and into the brain with its neural networks representing our thoughts in a hierarchy of neural correlates. A computational toolkit for optimal level prediction and control has been developed (Hoel and Levin 2020) which was conceptually extended to integrated information theory of consciousness *(Albantakis et al. 2019)*. The large gap between sciences and humanities discussed above exhibits itself in a series of small gaps connected to the emergence of informative higher scales. It has been suggested that the origin of life may be identified as a transition in causal structure and information flow (Walker 2014). Integrated information measures globally how much the causal mechanisms of a system reduce the uncertainty about the possible causes for a given state. A measure of "information flow" that accurately captures causal effects has been proposed (Ay and Polani 2008). The state of the art is presented in (Ay et al. 2022) where the link between information and complexity is discussed. Ay et al single out hierarchical systems and interlevel causation. Even further, (Rosas et al. 2020) reconcile conflicting views of emergence via an exact information-theoretic approach to identify causal emergence in multivariate data. As information becomes differentially richer one eventually needs complexity measures beyond $\{R^n\}$. One may define generalized metrices on these spaces (Pirró 2009) measuring information complexity on ever higher hierarchical levels of information. As one rises through hierarchies, information on higher scale is usually gained by coarse graining to arrive at an effective, nevertheless exact description, on the higher scale. It is repeated coarse graining of syntactically well-ordered information layers which eventually leads to semantic information in a process which I conjecture to be reminiscent of renormalization group flow leading to a universal classification scheme. Thus, we identify scientific disciplines and their corresponding data sets as dual universality classes of physical and epistemic structure formation, respectively. Above the semantic gap, we may call this process quantification of the qualitative by semantic metrics. Indeed, (Kolchinsky and Wolpert 2018) explored for the first time quantitative semantic concepts in Physics in their 2018 seminal paper entitled *Semantic information, autonomous agency and non-equilibrium statistical physics*. Their measures are numeric variants of entropy. Semantic information is identified with *'the information that a physical system has about its*

environment that is causally necessary for the system to maintain its own existence over time'.

FDOs are employed in these processes in two fundamental ways. For practical implementations of FDO technology, see accompanying abstract (Wittenburg et al. 2022). First, the FAIR principles (Wilkinson et al. 2016) ensure that unconnected pieces of data may be percolated into an integrated data space. Percolation creates the information density needed to feed AI-driven built up of semantic space. Without FDOs we wouldn't have the gravity for this to occur. Second, the very structure of FDOs, capable of symmetry preserving or breaking fusion events into composed entities, makes them homologous to mathematical categories. This will proof to be a powerful tool to unravel the nature of information via analyzing its topological structure algebraically, especially when considering our conjecture concerning universality, classes of information and their possible instantiations on vastly different length and time scales, in effect explaining analogous structure formation.

Keywords

categories, emergence, FDO, hierarchy, meaning, renormalization, semantic metric

Presenting author

Hans-Günther Döbereiner

Presented at

First International Conference on FAIR Digital Objects, presentation

Acknowledgements

References

- Albantakis L, Marshall W, Hoel E, Tononi G (2019) What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy 21 (5): 459.
- Anderson PW (1972) More is different: broken symmetry and the nature of the hierarchical structure of science. Science 177 (4047): 393-396.
- Ay N, Polani D (2008) Information flows in causal networks. Advances in complex systems 11 (01): 17-41.
- Ay N, Bertschinger N, Jost J, Olbrich E, Rauh J (2022) Information and Complexity, or: Where is the Information? In: Editor (Ed.) Lake Como School of Advanced Studies

Complexity and Emergence: Ideas, Methods, with special attention to Economics and Finance. Springer

- Bawden D, Robinson L (2020) Still minding the gap? Reflecting on transitions between concepts of information in varied domains. Information 11 (2): 71.
- Floridi L (2013) The philosophy of information. OUP Oxford
- Hoel E, Levin M (2020) Emergence of informative higher scales in biological systems: a computational toolkit for optimal prediction and control. Communicative & Integrative Biology 13 (1): 108-118.
- Kolchinsky A, Wolpert DH (2018) Semantic information, autonomous agency and non-equilibrium statistical physics. Interface focus 8 (6): 20180041.
- Pirró G (2009) A semantic similarity metric combining features and intrinsic information content. Data & Knowledge Engineering 68 (11): 1289-1308.
- Rosas FE, Mediano PA, Jensen HJ, Seth AK, Barrett AB, Carhart-Harris RL, Bor D (2020) Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. PLoS computational biology 16 (12): 1008289.
- Taddeo M, Floridi L (2005) Solving the symbol grounding problem: a critical review of fifteen years of research. Journal of Experimental & Theoretical Artificial Intelligence 17: 419-445.
- Walker SI (2014) Top-down causation and the rise of information in the emergence of life. Information 5 (3): 424-439.
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Mons B, (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3 (1): 1-9.
- Wittenburg P, Döbereiner HG, Weisgerber S, Morigi G (2022) FDO Project for Germany. In: Loo T (Ed.) First International Conference on FAIR Digital Objects. First International Conference on FAIR Digital Objects. Research Ideas and Outcomes https://doi.org/10.3897/rio.8.e95806