A Concise Taxonomy for Describing Data as an Art Material

Julie Freeman, Media & Arts Technologies, School of Electronic Engineering & Computer Science, Queen Mary University of London, U.K. Email: <j.freeman@qmul.ac.uk>. Geraint Wiggins. Email: <g.wiggins@qmul.ac.uk>.

Gavin Starks, Open Data Institute.

Mark Sandler. Email: <m.sandler@qmul.ac.uk>.

See <www.mitpressjournals.org/toc/leon/51/1> for supplemental files associated with this issue.

Submitted: 27 July 2016

Abstract

How can we describe data when used as an art material? As the number of artists using data in their work increases, so too must our ability to describe the material in a way that is understood by both specialist and general audiences alike. Based on a review of existing vocabularies, glossaries and taxonomies of data, we propose our own concise taxonomy. To conclude, we propose the adoption of this concise taxonomy by artists, critics and curators, and suggest that ongoing refinement of the taxonomy takes place through crowdsourced knowledge sharing on the Web.

Introduction

Data is no longer just in the domain of engineers and scientists. In fact it never was; designers and cartographers have been visualizing data for around 3,000 years [1]. Today, data are deeply embedded in all subject domains and within our daily lives. From the mundane to the specialist, whether 3D printing a kidney [2], doing your washing [3], scheduling a meeting, designing a city [4] or finding a partner [5], it takes some consideration to find an activity that does not involve data.

As electricity is pervasive in many societies, so too is digital data [6]: It has become a layer of essential infrastructure [7]. For clarity, we will use the word data in this paper to refer to digital (binary) data specifically: machine-readable, representing a set of distinct pieces of information (datum) in a particular structure and format that describe something.

So what do data mean to us? Again, like electricity, data are invisible yet necessary components in many of the systems that surround us. Enablers and disablers, data can inform decisions, help solve problems and provide insight. In their raw format they are sets of individual values that can be manipulated, reconfigured and transformed. This highly flexible, malleable substance is an ideal art material.

Artists need to understand any material they work with so that they can use it effectively to convey their ideas. The same applies to data, which are not usually framed as an art material. This lack of conceptualizing data as an art material has led us to notice that it does not often receive adequate depth of description when mentioned in interpretation texts supporting artworks. There is a difference between experiencing works that incorporate real-time data as opposed to historical data, or that depict a so-called truth garnered from a sample size of five participants versus 50,000 participants. To interpret the work fully, these differences should be made accessible to any audience.

In this paper, we consider why artists use data as a material and how data can be translated. Based on existing vocabularies used specifically within the arts, we propose a concise taxonomy for use in the description of data as an art material, designed for artists, curators, critics and associated general audiences. We conclude that although there are many taxonomies and vocabularies for cataloguing art, they are not easily adoptable tools in this context and that our concise taxonomy is more practical.

Through the definition of this working taxonomy we hope to encourage discourse around data as an art material and to enable comparison and critical review in a consistent manner. Our formal way of describing data can help reveal deeper understanding of the inclusion of data in artistic processes and enable us to gain insight into differences and similarities between artists in their conceptualization, approach and implementation of data in their work.

On Data

Data is a broad term that refers to collections of values that help us understand a phenomenon more deeply. It is used as a conceptual container for the reader to fill with facts and figures. Data are measurements of all kinds and can be used to generate more data. Euclid's book of propositions from around 300 B.C.E., *Data* [8], was written to "facilitate and promote the method of resolution or analysis," in other words to clarify what we can do with the data we have. His propositions (such as *if X then Y*) take givens (existing datum) and enable the deduction or inference of new data—a process we are very familiar with. *Datum* is a Latin term meaning "something given." In *The Data Revolution* [9] we read a quote by Jensen from 1950 that explains that really we should be referring to data as "capta," from the Latin *capere,* meaning "to take" [10]. It could be argued that we have lost the idea that data are a collection of things to be given, as opposed to taken [11].

Data (with their perception of benevolent evidence) can hold the promise of a new perspective, a digital version of the Overview Effect [12], and can be the foundation of many different outputs and experiences, such as graphic visualizations, artworks, animations, sound and music, narratives, tactile experiences, objects, scent, textiles and even personalized cosmetics.

Why Use Data as an Art Material?

As an art material, data has a great many attributes, including being low in cost (often free), widely available, easy to manipulate and abundant. It can even self-replicate. This variety and depth present a challenge to an artist who wishes to become fully proficient with a material they cannot handle directly. Although seemingly intangible, data can help illuminate and make sense of things we cannot see, feel or hear with our human senses. For an artist, it is a particular medium via which to be curious about the world.

There are many different ways data can be used in an artwork. For example, it can generate the essence of the work, allowing shapes and forms to be derived from the dataset itself [13]. It can be: used as a driver to generate dynamics [14]; mapped conscientiously to communicate a message; used to reveal patterns [15]; or misappropriated into artifice [16]. In *The Anti-Sublime Ideal in Data Art* [17], Manovich discusses mapping as the primary way of using data in art; this clearly identifies data as process but not data as material, framing it in computer science rather than fine art.

Given the ubiquity of digital technology, we argue that it is a legitimate material through which to reflect our lives and should be acknowledged as such. Data is at the heart of digital culture. Without its prevalence, the systems we rely on—from global finance through to personal communications—would fail. It is integral to governance, economics, social accord (and discord) and of course generation of, and access to, the arts.

Like the steam engine as a catalyst of the industrial revolution, and TV and radio bringing democratization to education, data is seen as the technology that will save us. How? By giving us the raw material with which to expose more knowledge than ever before, that is, to gain insight beyond expectations of the past. And as we instrument the world through sensors and mass-measurement, and data becomes infrastructure, the language we use to describe and to criticize it becomes paramount.

Translating Data

The impact of the delivery, type, properties and other characteristics of data on the creation and experience of an artwork is significant. If the work uses real-time data from a living source, what are the consequences of the death of the source? What does it suggest if the data transfer fails? If the data is anecdotal, or fabricated, is that made obvious? Does it need to be? Do preconceived ideas of data as evidence (real or not) reinforce the artist's intention? Does the intimacy of the work increase if the data is personal, or does it heighten discomfort? Is the temporal aspect of the work true to the data, or is the artist manipulating time?

The design and construction of the work can also affect how data is experienced. Obfuscation can take place within code through filters, randomness, subjective programming or biased algorithms. The aesthetic of the work can conceal or alter meaning derived from the data if it is over-bearing or has some strong characteristics. As Negroponte [18] says "the signature of the machine can be too strong," at the same time acknowledging the benefits of working with digital materials, in that "the process, not just the product, [can] be conveyed." These thoughts lead toward refinement of the way data art is described, and the level of detail about the core material, the data, that is included in those descriptions.

Existing Taxonomies and Vocabularies

Every taxonomy has a purpose—to elucidate information within a field, to define an index, to enable meaningful relationships to be made. Often they are created to work within existing higher-level ontologies, removing accidental duplication and furthering standardization. Cataloguing art is a wide and established field, particularly in media-based arts [19], which are in constant flux, as the materials change continually, even while part of a live work. Software and hardware redundancy rates are high and protocols and interfaces change and can become unusable very quickly [20]. In this oscillating culture, we can easily mislay important developments through an inability to log, capture and retrieve them. In addition, the lack of palpability of data elevates the need for careful metadata tagging and permanent linking, as without physical actuality, the retrieval of the work relies solely on future audiences being able to establish its digital existence.

During the development of this taxonomy, tagging and categorizing techniques in significant online artwork archives of net art, data art and media art were reviewed. Some visual taxonomies and relevant data glossaries were also studied. These included: the Getty Vocabularies (e.g. CDWA, AAT, CONA); the Dublin Core Metadata Initiative; Rhizome's Artbase [21]; the Archive of Digital Art; Turbulence.org; the Rose Goldsen Archive of New Media; Shneiderman's Data Type Taxonomy [22] and the updated version produced with Heer [23]; Visualising.org; Lima's Syntax of a New Language [24]; the U.S. White House's Project Open Data glossary [25]; and the U.K.'s <http://data.gov.uk> glossary.

It is evident from reviewing these archives, vocabularies and taxonomies, that there is a lack of consistency in the language used when describing data art and data visualization. Moreover, it is only the open data resources that mention the type, origin or delivery method of data. All the artwork archives fail to comprehensively describe data despite them being a core material in many works. It could be that not conceptualizing data *as a material* has led to the exclusion of comprehensive descriptors from the collections of terms referenced above.

A review of the substantial body of research on data visualization categorization and taxonomies that focus on the semiology, syntax and visual meaning of graphics (including Tufte, Bertin and von Engelhardt) is beyond the scope of this paper. The large number of technical data taxonomies, including the W3C data definitions and schema.org, are also beyond scope.

A Concise Taxonomy for Describing Data

Of living: Biological; Environmental

Of non-living: Object

Of social context: Commercial; Personal; Social; State

Of license: Closed; Open; Shared

Of time/space: Live; Real-time; Geospatial; Static; Temporal

Of type: Anecdata; Causal; Generated; Metadata; Processed; Retrieved; Streamed

Of disclosure: Anonymized; Identifiable; Unknown

Within an artwork, as opposed to a visualization, the viewer is allowed flexibility in translation. An artist may have the intention of provoking emotion or passing comment on a subject, but we cannot assume that it is the role of the artwork to convey a certain message due to the use of a particular dataset. This taxonomy is designed for artists, curators, critics and consumers of any art that incorporates data as a material. It is a descriptive set of terms, that is, it eschews some technical accuracy for classifications that are more commonly understood and easy to apply. To borrow from Guarino's ontology definitions [26], we have worked in a philosophical manner to create a set of words that form an informal conceptual system, which is that the terms underlie a more specific knowledge base (such as the CDWA). It is a challenge to represent all aspects of data in a uniform way; therefore, this taxonomy includes generic terms that guide the reader toward a richer understanding of the data and, perhaps, of why it is being used in the artwork.

We have aimed to create a concise taxonomy that enables data to be described in an objective way. Its purpose is not to describe subjective response of the viewer or listener; hence we have not included terms that can be applied to the affective descriptions of the experience of the work, such as *evocative* or *intimate*. We have also avoided terms that describe the aesthetic that the data yields in the artwork itself, such as *dynamic* or *abstract*. We acknowledge that while useful for categorizing and grouping art for some purposes, these more subjective terms are often personal and user-defined (by the artist, curator, audience, or critic), which makes a controlled vocabulary less effective and relevant.

The material (data) is examined from a number of perspectives—delivery method, how it emerged, format of existence, which system it represents, the source or origin, the license. In comparison, when considering a traditional art material, we may ask: where it was made, who made it, where it is from, what does it comprise, who owns it, how does it need to be stored, does it transform or degrade? Any number of the terms

in the taxonomy may be relevant to any one artwork, and it should be used with this in mind. For example, *Listening Post* by Mark Hansen and Ben Rubin [27] would be tagged *personal, social, live, real-time, temporal, retrieved, processed, anecdata.*

Definitions

Of Living: Biological—Data whose origin is directly linked to something that is alive. Data that occurs without conscious origin (i.e. not from a human typing). Often from sensors. Examples: (a) species migration reported by a sensor; (b) quantified self data such as output from a heart-rate monitor; (c) a birdcall.

Of Living: Environmental—Data whose origin is directly linked to the natural world. Often from sensors. Examples: (a) ocean temperature; (b) solar storm activity; (c) seed bank information.

Of Non-Living: Object—Data whose origin is a physical object or device. Object data is often generated for machine-to-machine communication; however, the Internet of Things will see a greater machine-to-(human) consumer communication. Examples: (a) a fridge's energy use; (b) a CCTV camera; (c) a smart watch.

Of Social Context: Commercial—Data produced by or about a corporate entity. Examples: (a) 10 years of financial information about a company; (b) the expiry date on a chocolate bar.

Of Social Context: Personal—Data produced by or about an individual. Certain types will have restricted access and some legal and technical protections. Others will be accessible by some, if not all, of the general public. Examples: (a) Google's search analysis profile of a non-anonymized individual's interests; (b) International travel logs held at border controls; (c) a recording of a private telephone conversation; (d) family photos publicly tagged on Flickr; (e) your social network feed.

Of Social Context: Social—Data produced by or about a social group or society. Examples: (a) global number of births each day; (b) voting preference in a London borough; (c) immigration figures.

Of Social Context: State—Data produced by or about a government or ruling authority. Examples: (a) the economy of the eurozone; (b) legislation documents.

Of License: Closed—Closed data is generally only accessible to people within an organization or to certain individuals. Examples: (a) company personnel files; (b) national security documents.

Of License: Open—Open data can be accessed, used and shared by anyone. Examples: (a) publicly funded research data; (b) earthquake monitoring data.

Of License: Shared—Shared data is data available to a specific group of people for a specific purpose. Examples: (a) the electoral register; (b) anonymized supermarket shopping patterns.

Of Time/Space: Live—Data that is, or was, captured in real time. The recording does not necessarily get played back at the same rate, or in the same moment. Examples: (a) a football match on TV; (b) animal tracking data.

Of Time/Space: Real-time—Data that is created, captured and disseminated in an immediate time-frame relative to the context of its use; it changes over time. Examples: (a) smart-meter reporting electricity usage every 30 seconds (real-time data acquisition with a relevant-time display); (b) feeds from sensors such as a webcam on a bird's nest, a GPS location of a mobile phone, or a humidity reading in a gallery space.

Of Time/Space: Geospatial—Data describing, is relevant to or is derived from a space or geographic area. Examples: (a) GPS coordinates from a cross-country walk; (b) the number of people visiting the Tate Modern art gallery; (c) the area of a baseball pitch; (d) longitude and latitude.

Of Time/Space: Static—Data in which the items do not change once created, but the dataset can grow over time. Includes historical datasets and archive indexes. Examples: (a) historical global population size; (b) a recording in the sound archive at the British Library.

Of Time/Space: Temporal—Data that is time-based in its nature, relevant to a specific time, or that may only exist for a short time period (transient). Examples: (a) the value of a kilogram of rice over time; (b) your date of birth; (c) the radio signals received from an exploding star.

Of Type: Anecdata—Anecdotal information gathered and presented as evidence. Anecdata is often not precisely measurable, has no reliable provenance, is hard to compare and/or cannot be unproven by the scientific method. Examples: (a) a collection of comments on a product website; (b) proverbs such as "Never look a gift horse in the mouth."

Of Type: Causal—Data in which it is (or is made) obvious to the observer what its origin is. Example: a vocal recording.

Of Type: Generated—Data created by a software program. Examples: (a) algorithmic music; (b) cellular automaton; (c) a model of a galaxy exploding.

Of Type: Metadata—Data about data. Data that describes information about other data. Examples: (a) the number of rows in a database; (b) the time and date a phone call was made.

Of Type: Processed—Data that has been calculated, altered or processed in some way. Examples: (a) a sonification of stock market figures; (b) aggregated statistics; (c) a colorful digital photograph reduced to black and white.

Of Type: Retrieved—Data made available on request by machine or user. Examples: (a) compilation of weather data from the past 24 hours as a single CSV file; (b) loan status of a library book.

Of Type: Streamed—The technical means of delivering real-time data as a continuous stream. The primary use-cases are where there is no requirement for data storage, or the datasets involved are too large to be manipulated in any other manner (the entire Twitter back catalogue). Examples: (a) real-time audio and video from a carnival procession; (b) on-demand replay of a film from 1960; (c) music playing from a digital radio.

Of Disclosure: Anonymized—Data that has had any identifiable information about a person, animal or thing removed. Examples: (a) CCTV camera footage containing people that has been blurred or obfuscated; (b) all bicycle hire users across a city with user IDs and names removed.

Of Disclosure: Identifiable—Data in which the direct source within it (person, animal or thing) can be identified. Examples: (a) a Facebook data export including friend names; (b) a set of mobile phone numbers with owner address details.

Of Disclosure: Unknown—Data that contains information about a person, animal or thing but in which it is not clear if it

is adequately anonymized. Examples: (a) a live Twitter feed containing some geolocated photos of people and animals; (b) a sound recording from a public space that includes ambient conversation.

Additional Dataset Parameters

There are aspects of data that are useful to explore in the process of understanding datasets that are not included in the taxonomy. These tend toward more technical descriptions and are used by archivists and preservation experts. The W3C Data on the *Web Best Practices Use Cases & Requirements Note* [28], recommends these elements are used for defining data: *domains*, *obligation/motivation*, *usage*, *quality*, *lineage*, *size*, *type/format*, *rate of change*, *data lifespan*, *potential audience*. We recommend considering the following, particularly for retrieval, maintenance and archival purposes of the artwork:

Accuracy: How exact are the individual data points (e.g. if it is real-time data is there latency to acknowledge).

Utility: Does the data have potential to provide utility by providing new content or insight, is this important to the work?

Provenance: Scientific datasets should be reproducible and should be collated from, or by, reliable sources. Any bias should be declared or detected.

Context: Does this dataset provide meaning through its relationships to other datasets (for comparative interest, for ratification)?

Relevancy: Are the data points relevant to each other, to someone or something (e.g. a machine)?

Accessibility: How and by whom can the dataset be accessed and used (licensing rights, availability, database rights) and is this reliable and future-proof?

Format: What is the structure and format (technical data structure and/or data definition, distribution)?

Dimensionality: How many dimensions are represented (e.g. a point against time, a number of parameters)?

Size: The order of magnitude of the number of data points, the sample size (e.g. 1 or 1 million). Often imprecisely referred to as large (big) data or small data.

A Note on Licensing—The taxonomy includes reference to *open, shared* and *closed* licenses. It is important to note that datasets are nearly all issued under some form of restriction. Even open datasets (available for free, to reuse, for any purpose) can have attribution requirements. Within artwork, which by default has copyright assigned to the artist, it is imperative that the use of a restricted material within it is acknowledged. Freeman's work *We Need Us (2014)* uses real-time open data from *zooniverse.org*. As the core material in the artwork is open, the ability for her to completely own the work outright is impossible, ownership must be reconsidered, therefore, the work has a series of different licenses that apply to various elements and uses of it [29].

A Note on Privacy and Anonymized Data—Much of the data used within artwork can be directly attributed to its source. Indeed, the revelation of the source often confers a large part of the meaning of the artwork. In the taxonomy, the *Of Disclosure* category includes a*nonymized*, *identifiable* and *unknown* tags. Whereas in other categories *unknown* is not specifically required, the declaration of using data in which it is not known whether it is anonymized is important.

Paolo Cirio & Alessandro Ludovico's work *Face to Facebook (2012)* [30] uses shared, easy to acquire, but unauthorized and identifiable scraped data to create a fictitious dating website. The controversy of the action would not exist if the data did not allow direct identification of real people.

The *disclosure* section of the taxonomy requires additional consideration on whether animals and certain objects have rights to privacy and whether re-identification possibilities through merging multiple datasets renders absolute anonymity possible.

Conclusion

The collaborative development and application of this taxonomy has highlighted that artists describe data in different ways making cross-referencing and comparison difficult, and that there is a lack of standardized terms to refer to. We note that the Getty vocabularies are complex, mainly for use by domain experts. Our taxonomy aims to be an accessible and adoptable way of categorizing data as an art material. We view the work as an accompaniment to Heer and Shneiderman's taxonomy of interactive dynamics for visual analysis, and as a potential addition to the Digital Art Archive.

The taxonomy is released on GitHub to encourage suggestions for ongoing improvement [31]. Through this public collaboration we aspire to contribute to the Project Open Data metadata schema, and perhaps the Getty vocabularies themselves. We also invite contributions to the data art database at <www.translatingdata.org>.

We conclude that the proposed taxonomy will be an aid to those archiving and cataloguing works in the future, but more importantly its lightweight nature should encourage use by practitioners, those new to the field of data art and others. In the words of Gillespie [32], we hope that it is "specific enough to mean something and vague enough to work across multiple areas for multiple audiences." The taxonomy prompts us to think about data as a material, and as such an essential component of any artwork that demands full disclosure.

Acknowledgments

Supported by RCUK Doctoral Training Centre EP/G03723X/1.

References and Notes

Based on a presentation given at VISAP'15, 19–30 October 2015, Chicago, IL, U.S.A. The IEEE VIS Arts Program (VISAP) showcases innovative artwork and research that explores the exciting and increasingly prominent intersections between art and visualization. The theme of VISAP'15 was Data Improvisations.

All URLs accessed 7 April 2016.

1. M. Friendly and D. J. Denis, "Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization" (2001), <www.datavis.ca/milestones>.

2. V. Mironov et al., "Organ Printing: Computer-Aided Jet-Based 3D Tissue Engineering," T*rends in Biotechnology* **21**, No. 4, 157–161 (April 2003).

3. R. M. Milasi, C. Lucas and B. N. Araabi, "Intelligent Modeling and Control of Washing Machine Using LLNF Modeling and Modified BELBIC," *Control and Automation* (2005).

4. A. Caragliu, C. DelBo and P. Nijkamp, "Smart Cities in Europe," *Journal of Urban Technology* **18**, No. 2, 65–82 (April 2011).

5. E.J. Finkel et al., "Online Dating: A Critical Analysis from the Perspective of Psychological Science," *Psychological Science in the Public Interest* **13**, No. 1, 3–66 (March 2012).

6. B. Sterling, "Digital Decay," in A. Depocas, J, Ippolito and C. Jones, eds., *Permanence Through Change: The Variable Media Approach* (New York/Montreal: Guggenheim Foundation/D. Langlois Foundation, July 2003).

7. ODI (2015), <http://theodi.org/who-owns-our-data-infrastructure>.

8. Euclides, *The Elements of Euclid*, R. Simson & W. Rutherford, trans. (1854).

9. R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (SAGE, August 2014).

10. H. Becker, "Science, Culture, and Society," *Philosophy of Science* **19**, No. 4 (October 1952) p. 273.

11. G. Greenwald, *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State* (Metropolitan Books, May 2014).

12. The Overview Effect: when the first images of Earth from outer space were broadcast, it fundamentally and irreversibly shifted our worldview: <http://www.overviewinstitute.org>.

13. Natalie Miebach (2011): <http://nathaliemiebach.com/gulf.html>.

14. YoHa (2011): <http://yoha.co.uk/invisible>.

15. Troika (2010): <http://youtu.be/DYp3hV0cM30>.

16. Benedict Groß (2012): <http://benedikt-gross.de/log/2012/02/metrography-london-tube-map-to-large-scale-collective-mental-map>.

17. L. Manovich, "The Anti-Sublime Ideal in Data Art" (2002), <http://virus.meetopia.net/pdf-ps_db/LManovich_data_art.pdf>.

18. N. Negroponte, *Being Digital* (New York: Random House, 1995).

19. B. Graham, "Taxonomies of New Media Art—Real World Namings," in *Museums and the Web 2005: Proceedings. Archives & Museum Informatics* (March 2005).

20. M.O. Ward, "A Taxonomy of Glyph Placement Strategies for Multidimensional Data Visualization," *Information Visualization* **1**, Nos. 3-4, 194–210 (December 2002).

21. R. Rinehart, "A System of Formal Notation for Scoring Works of Digital and Variable Media Art," in *REFRESH! The First International Conference on the Histories of Art, Science and Technology* (October 2005), doi:10002/307.

22. B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," in *Proceedings of IEEE Symposium on Visual Languages* (September 1996) pp. 336–343.

23. J. Heer and B. Shneiderman, "Interactive Dynamics for Visual Analysis," *Queue* **10**, No. 2 (February 2012) p. 30.

24. M. Lima, *Visual Complexity. Mapping Patterns of Information* (Princeton Architectural Press, August 2013).

25. Project Open Data <http://project-open-data.cio.gov/v1.1/schema/>.

26. N. Guarino, "Formal Ontology in Information Systems," in *Proceedings of the 1st International Conference* (Trento, Italy, Ios Press Inc, June 1998).

27. W. Modes, "Revisiting the Technical Achievements of *Listening Post* Ten Years On," in *The Journal of New Media & Culture* **9**, No. 1 (Winter 2014).

28. Data on the Web Best Practices Use Cases & Requirements, W3C Working Group Note, 24 February 2015: <https://www.w3.org/TR/dwbp-ucr/>.

29. Julie Freeman (2014): <http://weneedus.org/webpages/licence.htm>.

30. Paolo Cirio and Alessandro Ludovico (2012) <www.face-to-facebook.net/>.

31. Data Taxonomy (2014), <http://github.com/misslake/taxonomy-for-data-as-art-material/>.

32. T. Gillespie, "The Politics of Platforms," *New Media & Society* **12**, No. 3, 347–364 (2010).