A concise taxonomy for describing data as an art material.

FREEMAN, J; SANDLER, M; WIGGINS, G; STARKS, G; IEEE VIS

A Concise Taxonomy for Describing Data as an Art Material

Julie Freeman* Queen Mary University of London
Professor Geraint Wiggins† Queen Mary University of London
Gavin Starks‡ Open Data Institute
Professor Mark Sandler§ Queen Mary University of London

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. In this paper we review existing vocabularies, glossaries, and taxonomies of data, and propose our own concise taxonomy. We present a number of examples of how existing data art works are described, and demonstrate our taxonomy by applying it to these works. To conclude we propose the adoption of this concise taxonomy by artists, critics, and curators, and suggest that on-going refinement of the taxonomy takes place through crowd-sourced knowledge sharing on the web.

Keywords: Data, art, visualization, taxonomy

Index Terms: A.2 [Reference]: Taxonomies; E.m [Data]: Miscellaneous—Definitions

General Terms: Professional Communication: Taxonomies, Data Visualization, Data Art

1 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 [9]. Today, data are deeply embedded within all subject domains and into our daily lives. From the mundane to the specialist, whether 3D printing a kidney [21], doing your washing [20], scheduling a meeting, designing a city [4, 5], or finding a partner [7], 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[1]: it has become another layer of essential infrastructure[2]. 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 which describe something.

So what do data mean to us? Again, like electricity, data are invisible yet necessary components in many of the systems which 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 which 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 them effectively to convey their ideas. The same applies to data, which are not usually framed as an art material. This lack of conceptualising 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 which incorporate real-time data as opposed to historical data, or which 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. We then look at existing vocabularies used specifically within the arts. Based on our initial findings, 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. Through examining how a number of artists refer to data when describing their work, we note whether or not our taxonomy terms are synonymous with the language in the practitioners descriptions. 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.

1.1 Motivation

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 work will assist in revealing a deeper understanding of the inclusion of data in the artistic process, and help us gain insight into differences and similarities between artists in their conceptualisation, approach, and implementation of data in their work. In addition, a formal way of describing data is important as it becomes prolific as an artistic material and as data types and tools evolve.

2 ON DATA

Data is a broad term that refers to collections of values which 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 300BCE, *Data* [6], 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[3]) and enable the deduction or inference of new data—a process we are very familiar with. How would Euclid respond to today's data-driven world?

Data (with their perception of benevolent evidence) can hold the promise of a previously unseen overview[4] from a different perspective, and can be the foundation of many different outputs and expe-

*e-mail: j.freeman@qmul.ac.uk
†e-mail: g.wiggins@qmul.ac.uk
‡e-mail: gavin@theodi.org
§e-mail: m.sandler@qmul.ac.uk

[1] Bruce Sterling, science fiction author, refers to digital data, as opposed to general data, as data which is generated with, or stored on, a computer. [27]

[2] `http://theodi.org/who-owns-our-data-infrastructure?` Accessed 24 June 2015.

[3] Datum is a Latin term meaning 'something given'. In *The Data Revolution* [16] we read a quote by Jensen from 1950 (originally cited in [2]) which explains that really we should be referring to data as 'catpta' from the Latin 'capere' meaning 'to take'. Have we lost the idea that data are a collection of things to be given, as opposed to taken [13]?

[4] The *Overview Effect* is a photographic analogy—when the first images of the earth from outer space were broadcast it fundamentally and irreversibly shifted our world view. `http://www.overviewinstitute.org/about-us/declaration-of-vision-and-principles`. Accessed 2 June 2015

riences, such as graphic visualisations, artworks, animations, sound and music, narratives, tactile experiences, objects, scent, and textiles, and even personalised cosmetics[5].

2.1 Why Use Data as an Art Material?

As an art material[6] 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.[7] It can be: used as a driver to generate dynamics[8]; mapped conscientiously to communicate a message; used to reveal patterns;[9] or misappropriated into artifice.[10] In *The Anti-Sublime Ideal in Data Art*, Manovich [19] 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 the current 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 democratisation 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.

2.2 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? Whether the answers to these questions are of importance to the way in which the work is interpreted is up to the artist, but for comprehensive critical review, they are essential. Jer Thorpe[11], artist and educator, author of *Beautiful Visualisation* and *Data Flow 2*, comments:

> "The biggest failure of data art, in my opinion, is in neglecting to address the individual character of a data set. [...] Almost any data set you find has some specific character that could and should be addressed in a visualization—and certainly in a data art project." [3]

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 [23] says "the signature of the machine can be too strong", at the same time acknowledging that the benefits of working with digital materials is that "the process, not just the product, [can] be conveyed". These thoughts point us 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.

3 Existing Taxonomies & 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 standardisation.

Cataloguing art is a wide and established field, provoking ongoing debate [11, 12]. Media-based arts are in constant flux as the materials change continually, even whilst part of a live work. Software and hardware redundancy rates are high, protocols and interfaces change and can become unusable very quickly [28]. 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. As an example, unlike finding an Old Master in an attic, a seminal piece of net art from 1996 could easily become redundant: stored on a powered-down server, never to be seen again.

Following is a summary of some significant on-line artwork archives of net art, data art, and other media art. Within these tagging and categorising techniques are reviewed. Then follows mention of a small sample of visual taxonomies. A review of the substantial body of research on data visualisation categorisation and taxonomies that focus on the semiology, syntax and visual meaning of graphics is beyond the scope of this paper.[12] Two highly relevant and recent data glossaries are highlighted, however, there are a large number of technical data taxonomies, including the W3C definitions, which, again, are beyond the of scope of this paper.

3.1 The Getty Art & Architecture Thesaurus (AAT)

The AAT is a comprehensive structured vocabulary for describing and cataloguing art, architecture, and cultural heritage.[13] The vocabularies have been released as linked open data (LOD) which the authors believe will have "a truly transformative effect on the discipline of art history in general, and on Digital Art History in particular." The AAT is aimed at domain experts—curators, taxonomists, archivists. Viewable as a semantic hierarchy, in JSON, RDF, and

[5] Data-driven beauty http://www.fitnyc.edu/files/pdfs/Digital_Analytics.pdf. Accessed 22 June 2015.

[6] Without getting overly semantic or physical, we chose to use the word material over medium as the word medium has greater association with transference, data as a method to communicate, as opposed to an integral part. In addition, digital data is formed of electrons which are classed as matter. In this physical sense, data as material is valid. If it all ends up as photons that's another story.

[7] http://nathaliemiebach.com/gulf.html. Accessed 18 June 2015.

[8] http://yoha.co.uk/invisible. Accessed 18 June 2015.

[9] https://youtu.be/DYp3hV0cM30. Accessed 21 June 2015.

[10] http://benedikt-gross.de/log/2012/02/metrography-london-tube-map-to-large-scale-collective-mental-map. Accessed 20 June 2015.

[11] http://blog.blprnt.com/about. Accessed 23 June 2015.

[12] See work by Edward A. Tutfe, Jacques Bertin, and Jörg von Engelhardt.

[13] http://www.getty.edu/research/tools/vocabularies/aat/. Accessed 21 June 2015.

other ontological views, it is a vast set of generic terms that encompass a huge range of topics. It is not an easy reference to navigate and, although there are many data categories, such as *metadata, data processing, data loggers,* and so on, we could not locate a set of terms for describing data as a material within an artwork.

3.2 New Media Art Databases

The Rhizome ArtBase[14], established in 1999, is one of the most developed and medium-aware of the databases in the media arts field. The ArtBase is not just an index or catalogue, it archives the artworks and attempts to preserve and update them as technologies progress. Rinehart [25] completed extensive work summarising existing notations for creating a database and scoring system for artworks, underpinning this he previously described a metadata system that would work using existing schema elements from Dublin Core and the Categories for Description of Works of Art (CDWA-lite) [24]. The schema is viewable in the appendix of *Digital Preservation Practices and the Rhizome ArtBase* [8]. Tags *data* and *database* are used for the works, but no comprehensive data descriptions are present, leaving scope for a data category to be added to the metadata schema.

The Digital Art Archive[15] is a community-led catalogue. It uses an interesting taxonomy: under the section Aesthetics are listed *processual, sublime, vicinity*, and *inebriation (frenzy)*—words that are perhaps included at the point of submission by the artist. Under the technology section more standard terms are listed: *display, interface,* and *software,* with more detailed tags one level below. We suggest that *Data* and the proposed sub-categories are added to this section. This archive, due to the ability for its knowledgeable community to submit content, has a good chance of becoming a high quality reference site for, amongst other genres, Data Art.

At the forefront of cataloguing networked art, *Turbulence.org* has commissioned, exhibited, and archived net art for over 19 years. Their archival system is based on blog author metatags. Aside from the tag cloud, there is no obvious taxonomy published. An analysis of the existing metadata would be of great interest. The Turbulence website is currently a static archive on the Rose Goldsen Archive of New Media Art[16] at Cornell University Library. The Rose Goldsen archive is a resource for CD-ROM, DVDs, and ,increasingly, online artworks. It is a slowly developing work in progress, and does not appear to use standardised cataloguing or tagging techniques at the present time.

3.3 Visual Design Taxonomies

Shneiderman [26] sets about his data type taxonomy by listing seven tasks performed on data from a user perspective: *overview, zoom, filter, details-on-demand, relate, history, extract*. This is followed by seven data types: *1-dimensional, 2-dimensional, 3-dimensional, temporal, multi-dimensional, tree, network*. The taxonomy has some useful and usable terms. The updated work by Heer and Shneiderman [15] greatly improves on this, and focuses more deeply on the process and user, whilst dropping the data type terms—they are only loosely referred to as examples *multivariate, geospatial, textual, temporal, networked*. Both these works are a key influence on the taxonomy that follows.

Visualizing.org[17] has a simple taxonomy of visualisation techniques which is based on diagrammatic methods such as maps, charts, networks. In *The Book of Trees* [18], Lima explores the evolution of visualisation, using a visual tree language to categorise the techniques used, and in *Visual Complexity* [17] he creates his *Syntax of a New Language*, also a visual reference. These visual taxonomies are aimed at designers and complement the more descriptive work of Shneiderman and our proposed taxonomy.

3.4 Data Glossaries

A number of comprehensive definitions of data have come from the relatively new (since around 2007 [1]) open data movement. Both the US and UK governments have on-line glossaries which provide some reference for how to label data. It is possible that these glossaries[18] would benefit from merging. The US glossary has a metadata section which is a condensed version of the Project Open Data metadata schema.[19] Project Open Data (founded by the White House) has a comprehensive open data glossary and detailed metadata schema aimed at anyone interested in open data. With its roots in the Dublin Core Library, it is a useful and relevant resource.

3.5 Summary

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 visualisation. Moreover, it is only the open data resources which mention of the type, origin, or delivery method of data. All of the artwork archives fail to comprehensively describe data despite them being a core material in many works. It could be that not conceptualising data *as a material* has led to the exclusion of comprehensive descriptors from the collections of terms referenced above.

4 A Concise Taxonomy for Describing Data

Table 1: A Concise Taxonomy for Describing Data as an Art Material

Of living	Biological; Environmental
Of non-living	Object
Of social context	Commercial; Personal; Social; State
Of licence	Closed; Open; Shared
Of time or space	Live; Real-time; Geospatial; Static; Temporal
Of type	Anecdata; Causal; Generated; Metadata; Processed; Retrieved; Streamed
Of disclosure	Anonymised; Identifiable; Unknown

The taxonomy (see table 1) was first compiled in 2012 by Julie Freeman during her Masters studies at Queen Mary University of London (unpublished), together with Gavin Starks, from the Open Data Institute. It has since been refined through a mixture of informal qualitative research, including an examination of existing data art work descriptions, and from direct experience of working with data artists. In addition, Freeman has worked with data as a material in her art practise for many years.

Within an artwork, as opposed to a visualisation, 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 which 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 [14], we have worked in a philosophical manner to create a set of words that form

[14] http://rhizome.org/artbase/. Accessed 21 June 2015.

[15] https://www.digitalartarchive.at/nc/database/database-info/keywords.html. Accessed 21 June 2015.

[16] http://goldsen.library.cornell.edu/. Accessed 20 June 2015.

[17] http://www.visualizing.org/stories/taxonomy-data-visualization. Accessed 21 June 2015.

[18] https://www.data.gov/glossary and http://data.gov.uk/glossary. Both accessed 23 June 2015.

[19] https://project-open-data.cio.gov/v1.1/schema. Accessed 16 June 2015.

an informal conceptual system, which is that the terms underlie a more specific knowledge base (such as the Getty Art & Architecture vocabulary and the Project Open Data metadata schema). It is a challenge to represent all aspects of data in a uniform way, therefore this taxonomy includes generic terms which guide the reader toward a richer understanding of the data, and perhaps why it is being used in the artwork.

We have aimed to create a concise taxonomy which 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 whilst useful for categorising 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 is it from, what does it comprise of, 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 [22] would be tagged *personal, social, live, real-time, temporal, retrieved, processed, anecdata.*

4.1 Definitions

This section contains descriptions and examples for each term in the taxonomy as introduced above in table 1.

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 bird-call.

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.

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.

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.

Social Context: Personal Data produced by or about an individual. Certain types will have restricted access, and some legal and technical protections. Other will be accessible by some, if not all, of the general public. Examples: a) Google's search analysis profile of a non-anonymised 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.

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.

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

Licence: Closed Closed data is generally only accessible to people within an organisation or to certain individuals. Examples: a) company personnel files; b) national security documents.

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

Licence: Shared Shared data is data available to a specific group of people for a specific purpose. Examples: a) the electoral register; b) anonymised supermarket shopping patterns.

Time/Space: Live Data which 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.

Time/Space: Real-time Data that is created, captured and disseminated in an immediate[20] 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 birds nest, a GPS location of a mobile phone, or a humidity reading in an gallery space.

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.

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.

Time/Space: Temporal Data which is time-based in its nature, relevant to a specific time, or which 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.

Type: Anecdata Anecdotal information gathered and then 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".

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

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

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

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

[20] Immediate is approximate, and assumes some minimal system latency. In a video stream end-to-end latency would be due to encoding, post-processing, network processing, buffering, decoding, and pre-processing (see `http://www.cast-inc.com/blog/white-paper-understanding-and-reducing-latency-in-video-compression-systems`). Accessed 20 June 2015. For acceptable latency times that ensure user engagement varies, refer to Jakob Nielsen's work.

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) availability status of a library book.

Type: Streamed The technical means of delivering real-time data as a contiguous stream. The primary use-cases are where there is no requirement for data storage, or that the data-sets 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.

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

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.

Disclosure: Unknown Data which contains information about a person, animal, or thing but in which it is not clear if it is adequately anonymised. 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.

4.2 Additional Dataset Parameters

There are aspects of data that are useful to explore in the process of understanding datasets which 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*[21], 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 (see table 2).

4.3 A Note on Licensing

The taxonomy includes reference to *open, shared* and *closed* licences. 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 recent work, *We Need Us*, 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. Therefore, the work has a series of different licences that apply to various elements and uses of the work.[22] Using certain types of data as an art material requires us to reconsider ownership of the work.

4.4 A Note on Privacy and Anonymised 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 *anonymised, 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 anonymised is important.

[21] `http://www.w3.org/TR/dwbp-ucr`. Accessed 20 June 2015.

[22] Licence details `http://weneedus.org/webpages/licence.htm` and artwork `http://www.weneedus.org`. Both accessed 22 June 2015.

Table 2: Additional Dataset Parameters

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, others 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.

Endless War[23] by YoHa (with Matthew Fuller) uses the wikileaked Afghan War Diaries as it's core material. This data contains "...over 91,000 (15,000 withheld) reports covering the war in Afghanistan from 2004 to 2010. The reports were written by soldiers and intelligence officers...". The work takes a month to visualise the data set presenting the potential to reveal closed, confidential, but identifiable data—an aspect of the work that gives it gravitas and relevance.

Paolo Cirio's work *Face to Facebook*[24] uses shared, easy to acquire, but unauthorised and identifiable scraped data to create a fictitious dating website. The controversy of the action would not exist if the data did not link us directly to real people. Further, Cirio sources hard-to-acquire identifiable data in *Overexposed*[25], a work which publicly displays billboard sized photos of unauthorised high-ranking U.S. intelligence officials. Taking officials who hope to remain anonymous and putting them into public view uses the power of anonymity to make the work anything other than a series of photos on walls.

The *disclosure* section of the taxonomy requires more thought, including consideration on whether animals and certain objects have rights to privacy, and whether re-dentification possibilities through merging multiple datasets renders absolute anonymity possible.[26]

[23] `http://yoha.co.uk/node/761`. Accessed 20 August 2015.

[24] `http://www.paolocirio.net/work/face-to-facebook`. Accessed 20 August 2015.

[25] `http://www.paolocirio.net/work/hd-stencils/overexposed`. Accessed 20 August 2015.

[26] The UK government have produced an Anonymisation Code of Practice for personal data: `https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf`. Also see `http://ukanon.net`. Both accessed 20 August 2015.

5 Using the Taxonomy

During the development of this taxonomy a small database[27] of data art has been maintained as a resource for applying the terms to existing works. The database contains works which sit on a spectrum spanning fine art to visualisation tools to technical display, although many pieces are hard to pinpoint precisely on this scale. From this database we have selected five exemplars, choosing works from artists with varying levels of experience, technical expertise, and exposure. Visual, sonic, installations, screen and non-screen based works are included. We aimed to select both known and possibly unknown works in an attempt to represent a broad range of practitioners. For more case studies please refer to the database—it contains around 40 works that have been tagged using our taxonomy. Through a review of both the short descriptions and longer texts (where available), we look at how the artist refers to data, and then apply terms from our vocabulary.

5.1 dataMorphose (2009) by Christiane Keller

Figure 1: dataMorphose by Christiane Keller

This work (see figure 1) is a physical representation of data. It is an example in which data is fundamental to the work conceptually and aesthetically. Summarising the work, the artist states:

> ''dataMorphose is an interactive installation which projects **data** into real space and visualizes it three-dimensionally. Information is represented by spanned and moving sails directly in the room. Thus **abstract and virtual data** becomes real and tangible. As the user takes new positions and perspectives, he can experience a completely novel and sensual perception of **data**. Three spatial displays visualize **statistical data, web activities** and the **current time**. The **coding and procurement of data** is visualized by the tension of the canvas, the pace of movement, the position of the canvas and the change of their shape."
> From `http://www.christianekeller.de/datamorphose`. Accessed 20 June 2015. Emphasis added.

Keller refers to generic "data" a number of times, "information", and "abstract and virtual data". She goes some way to explain the data in more detail in the phrase "statistical data, web activities and the current time". Applying our taxonomy we try to describe the three datasets in more detail. Current time is a straight forward concept—*real-time, live, retrieved, temporal* and *geospatial* (it is always associated with a time zone). It is also *open* (public domain). For "web activities" we could make the assumption that the data is *real-time, retrieved, shared, metadata* based on the sources cited (Google Trends, Google Insights for Search). As we do not know the origin, type, and context, we are left unable to categorise the data further (*unknown*). The term "statistical data" also leaves us without clarity. We can assume only inclusion of *processed* data (that the values are the output from some statistical analysis). In the longer description the artist refers simply to "values and parameters" but without any details. Keller also refers to the "coding and procurement of data" being visualized in the work, however, despite reading the extended description it is hard to ascertain what this means. We conclude that the work seeks to demonstrate kinetic potential of data through physical form with the content of the data itself of little relevance to this.

[27] `http://translatingdata.org`. Accessed 22 August 2015.

5.2 A Conversation Between Trees (2012) by Active Ingredient

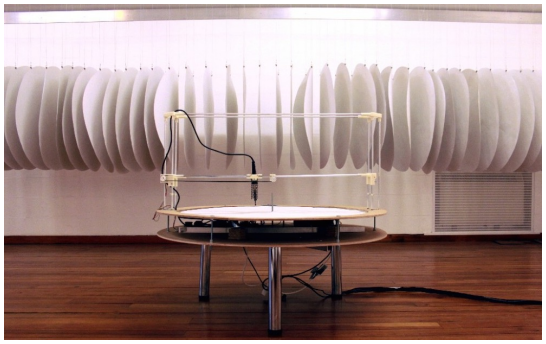

Figure 2: A Conversation Between Trees by Active Ingredient. Partial installation view

A Conversation Between Trees (see figure 2) uses a number of data sources, each bringing something to the experience—real-time sensor data mixed with scientific records. It is the meeting of the past and the present that reinforces the artistic concept about a need to act now for climate change.

> "[*A Conversation Between Trees is*] an exhibition that generates clues of the climate and history of our forests in the UK and Brazil... a screen flickers and glows with a dynamic 3D visualisation of **changes in temperature, humidity, light, decibels, colour and CO2 collected from trees** in both forests. Hanging from the ceiling is a full set of **global C02 data** scorched into circular sheets. Each sheet shows **a year** of changes in **C02 levels** in the Earth's atmosphere as a scorched ring. The prints will show a steady **annual increase recorded over the last 53 years** since scientific records began. A box attached to tree branches in both locations, contains sensors that sense **levels of temperature, humidity, C02, light, colour and sound levels**, which is **sent live** to the gallery via the internet."
> From `http://hello-tree.com/exhibition`. Accessed 20 June 2015. Emphasis added.

In the longer descriptive text the artists refer to *environmental, temporal, real-time, streamed, static, processed* and *metadata* albeit in less direct terms. Some references to 'live' are used to mean *real-time*. There are elements of assumption here too, that the reader will understand that data is informing change in the artwork, rather than the heat or humidity themselves, this enables us to add the *identifiable* tag. The description of the project and the variety of data used as a material, could benefit from using standardised keywords from the taxonomy to enable clearer categorisation and comparison.

5.3 Mori (1999) by Ken Goldberg, Randall Packer, Gregory Kuhn, and Wojciech Matusik

Mori (see figure 3) is an early example of data art. It began as a minimal visual on-line work in 1998 and was developed into an

Figure 3: Mori: an internet-based earthwork by Ken Goldberg, Randall Packer, Gregory Kuhn, and Wojciech Matusik (1999). Photo taken at at ICC Tokyo, November 1999, by Takasi Otaka

installation and on-line audio work 1999 and 2003, using the same source of live data.

> "*Mori* engages the earth as a **living** medium. **Minute movements** of the Hayward Fault in California are detected by a **seismograph**, converted to digital signals, and **transmitted continuously** via the Internet to the installation. Inside the entry curtain, visitors follow a fiber optic cable to the center of the resonating enclosure, where a portal through the floor frames the installation's focal point. The **live seismic data stream** drives an embedded visual display and immersive low-frequency sounds, which echo the unpredictable **fluctuations** of the earth's movement."
> From `http://goldberg.berkeley.edu/art/mori`. Accessed 20 August 2015. Emphasis added.

The work employs data that is *live, real-time, environmental, shared* (this may well be classified as *open* now), *streamed, identifiable, geospatial*, and *temporal*. The artists description is comprehensive, clearly reflecting on the importance of the data to convey the aliveness of the work. A recent (2013) development of Mori is an on-line visualisation, called *Bloom* produced in collaboration with Sanjay Krishnan, Fernanda Viegas, and Martin Wattenberg.[28]

5.4 The Shaping Grows (2012) by Semiconductor

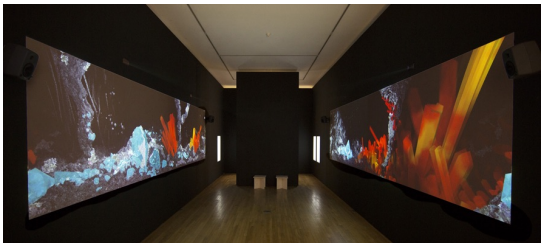

Figure 4: The Shaping Grows by Semiconductor. Image: David Levene

This sonic and visual installation (see figure 4) uses real world data to influence dynamic animation. The artists appear to have sought to generate a sense of aliveness within the work, even though the data is not in real-time.

> "*The Shaping Grows* is a computer **generated** animation of a subterranean cavern, brought to life through **seismic data**... The animation spans **multiple time frames** condensing geological events and **processes** through **time-lapse techniques**... crystals can become consumed by larger formations or play host to wildly different structures, as physical conditions **change over time** and favour certain elemental and chemical reactions... objects store the memory of their making and can be read to learn the story of their evolution and the conditions in which they grew. Semiconductor have collected **seismic data** of recent earthquake activity from around the world and **converted** it into sound. This directly animates and controls the formations and provides a sound-scape of the Earth in a state of flux."
> From `http://semiconductorfilms.com/art/the-shaping-grows`. Accessed 20 June 2015. Emphasis added.

This work has a multi-layered approach to data, and is on the whole well described. The work contains *static, environmental, geospatial, temporal, processed, identifiable*, and *generative* data. These are referred to in the description explicitly—processed and generated—and obliquely as "time-lapse techniques" and "multiple time frames". The minimal description reads: **03.00 minute loop, 4 channel HD + 4 channel audio**, and yet the core material in the work which "directly animates and controls" it is seismic *(environmental)* data. The general description provides a good sense of the data in the work demonstrating the artists comprehension of the material. However, we suggest that even in the minimal materials description data could be acknowledged.

5.5 The Live Wire (1995) by Natalie Jeremijenko

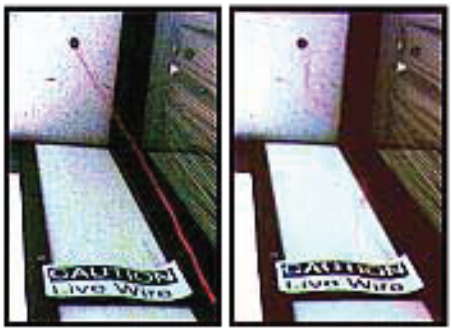

Figure 5: The Live Wire (1995)

A simple yet effective artwork that relied on data as a core material, *The Live Wire* (see figure 5), was developed by Natalie Jeremijenko whilst she was artist-in-residence at Xerox PARC in 1995.

> "*The Live Wire* is a 3D, **real-time** network traffic indicator. It is actually a material manifestation of cyberspace. Plugging into a local area network, it wiggles proportionally to the **amount of traffic on the net**. With each **data package** it convulses and sets up standing waves. It is placed in the spectacularly banal office environment of the glamorous Xerox Park Computer Science Lab, the place where WYSIWYG, Macintosh interface, ethernet and many other things were invented [...] Live Wire could be another graph on your computer screen, a **real-time 3D rendering of network traffic**, [...] But instead it is in the periphery, in the shared physical space."
> Edited from `http://tech90s.walkerart.org/`

[28] `http://goldberg.berkeley.edu/art/Bloom`. Accessed 20 August 2015.

nj/transcript/nj_04.html. Accessed 20 August 2015. Emphasis added.

The work is a fundamental representation of a data stream. It contains *real-time, live, closed, temporal, anonymised, streamed* data, all of which are easy to ascertain from the description. We have chosen to use the *closed* tag as it is unlikely that the local area network traffic information would be made available to anyone outside of Xerox PARC.

6 Conclusion

The concise taxonomy for describing data used as an art material has been developed collaboratively and applied to a sample of artworks as a method of testing its usability and relevance. This process has highlighted that artists describe data in different ways making cross-referencing and comparison difficult, and that there is a lack of standardised terms to refer to.

We note that the Getty vocabularies are complex, and are mainly used by domain experts. The aim of our taxonomy is to create an accessible, and adoptable, way of categorising data as an art material. We view the work as a neighbourly accompaniment to Heer and Shneiderman's taxonomy of interactive dynamics for visual analysis, and as a potential addition to the Digital Art Archive.

Current development work on the taxonomy includes public and targeted surveys, and its release on GitHub (see https://github.com/misslake/taxonomy-for-data-as-art-material/) to encourage a comments and suggestions for on-going improvement. 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 found at http://translatingdata.org, which, in time, will be available as open data.

We conclude that the proposed taxonomy will be an aid to those archiving and cataloguing works in the future, but more importantly its light-weight nature should encourage use by practitioners, those new to the field of data art, and beyond[29]. In the words of Gillespie [10], 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 which demands full disclosure.

Acknowledgements

This work is supported by the Media and Arts Technology programme, EPSRC Doctoral Training Centre EP/G03723X/1. The authors wish to thank our anonymous reviewers for their valuable contributions.

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[2] H. Becker. Science, Culture, and Society. *Philosophy of Science*, 19(4):273, Oct. 1952.

[3] K. Buford. Data Art vs. Visualization? The Distinction is Unproductive, says Artist Jer Thorpe. *http://siliconangle.com/blog/2012/08/22/data-art-vs-visualization-the-distinction-is-unproductive-says-artist-jer-thorp-qa/*. Aug. 2012.

[4] A. Caragliu, C. Del Bo, and P. Nijkamp. Smart Cities in Europe. *Journal of Urban Technology*, 18(2):65–82, Apr. 2011.

[5] H. Chourabi, T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. A. Pardo, and H. J. Scholl. Understanding Smart Cities: An Integrative Framework. In *2012 45th Hawaii International Conference on System Sciences (HICSS)*, pages 2289–2297. IEEE.

[6] Euclides, R. Simson, and W. Rutherford. The Elements of Euclid (translation), 1854.

[7] E. J. Finkel, P. W. Eastwick, B. R. Karney, H. T. Reis, and S. Sprecher. Online Dating: A Critical Analysis From the Perspective of Psychological Science. *Psychological Science in the Public Interest*, 13(1):3–66, Mar. 2012.

[8] B. Fino-Radin. Digital Preservation Practices and the Rhizome ArtBase. *http://media.rhizome.org/artbase/documents/Digital-Preservation-Practices-and-the-Rhizome-ArtBase.pdf*, July 2011.

[9] M. Friendly and D. J. Denis. Milestones in the history of thematic cartography, statistical graphics, and data visualization. *http://www.datavis.ca/milestones*. 2001.

[10] T. Gillespie. The politics of platforms. *New Media & Society*, 12(3):347–364, 2010.

[11] B. Graham. Taxonomies Of New Media Art – Real World Namings. In *Museums and the Web 2005: Proceedings*. Archives & Museum Informatics, Mar. 2005.

[12] B. Graham. Edits from a CRUMB discussion list theme. *Curating Immateriality: The Work of the Curator in the . . .*, 2006.

[13] G. Greenwald. *No Place to Hide*. Edward Snowden, the NSA, and the U.S. Surveillance State. Metropolitan Books, May 2014.

[14] N. Guarino. *Formal Ontology in Information Systems*. Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy. Ios PressInc, 1998.

[15] J. Heer and B. Shneiderman. Interactive Dynamics for Visual Analysis. *Queue*, 10(2):30, Feb. 2012.

[16] R. Kitchin. *The Data Revolution*. Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE, Aug. 2014.

[17] M. Lima. *Visual Complexity*. Mapping Patterns of Information. Princeton Architectural Press, Aug. 2013.

[18] M. Lima. *The Book of Trees*. Visualizing Branches of Knowledge. Princeton Architectural Press, Apr. 2014.

[19] L. Manovich. The Anti-Sublime Ideal in Data Art. *http://www.manovich.net/DOCS/data_art.doc*, Aug. 2002.

[20] 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.

[21] V. Mironov, T. Boland, T. Trusk, G. Forgacs, and R. R. Markwald. Organ printing: computer-aided jet-based 3D tissue engineering. *Trends in Biotechnology*, 21(4):157–161, Apr. 2003.

[22] W. Modes. Revisiting the technical achievements of listening post ten years on. *The Journal of New Media & Culture*, 9(1), Winter 2014.

[23] N. Negroponte. *Being Digital*. Vintage, 1996.

[24] R. Rinehart. Preserving the Rhizome ArtBase. *http://media.rhizome.org/artbase/documents/Preserving-the-Rhizome-ArtBase.pdf*, July 2002.

[25] R. Rinehart. A System of Formal Notation for Scoring Works of Digital and Variable Media Art. *http://www.bampfa.berkeley.edu/about/formalnotation.pdf*, June 2005.

[26] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343, Sept. 1996.

[27] B. Sterling. Digital Decay in *Permanence Through Change: The Variable Media Approach*, July 2003.

[28] M. O. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210, Dec. 2002.

[29] As citizens become more familiar with data through the growing interest in the Internet of Things, this taxonomy is relevant here too.