Theory of Dataphoric Space

Emerging Research Forum (ERF)

Michael J. Pritchard
Dakota State University
michael.pritchard@trojans.dsu.edu

Cherie Noteboom
Dakota State University
cherie.noteboom@dsu.edu

Abstract
A dataphore is akin to a species of animal within a biological biome. Humans facilitate the movement of information between dataphores by taking knowledge and insights from around us and transposing that knowledge into other dataphoric forms. Like the relationship between bees and flowers - information is the pollen that is exchanged between us. Humans cultivate and mediate the flow of data within dataphoric space. Our role as the predominant content mediator is not a singular role as we have created artificial dataphores to assist in the cultivation of data (e.g., data crawlers, data recognizers, data scrapers, data cleansers). Like data contained within a genetic sequence; its topology drives its expression. From tiny cellular data forms up to complex of dataphores, the topology of the data contained within a dataphora also drives its expression. We explore our creation of the dataphoric ontology and the associated research constructs that operate within dataphoric space.

Keywords
dataphoric space, dataphora, dataphore, entropy, ontology, semantics

Introduction
Information systems research ontologies have been described as lacking academic rigor (Alavi and Leidner, 2001). An important discussion began in 1998 when Boisot introduced the concept of Information Space. Boisot’s ontology describes knowledge types and its evolvement within Information Space. For example, knowledge can be converted from "proprietary knowledge" to "textbook knowledge" (Boisot, 1998). Alavi and Leidner would propose a more rigorous treatment of the ontological discussion (Alavi and Leidner, 2001). They describe the following knowledge types: tacit, explicit, individual, social, declarative, procedural, causal, conditional, relational and pragmatic. Both taxonomies have common ground (i.e., "Common Sense" is essentially "Tacit and Individual"). Each attempted to show the movement of information within information spaces, "If knowledge is a process, then the implied knowledge management focus is on knowledge flow and the processes of creation, sharing, and distribution of knowledge." (Alavi and Leidner, 2001). Yet, something quantitative to information systems research ontologies is missing. While researchers have made great progress in furthering the ontological discussion, inconsistencies remain. According to Alavi and Leidner, tacit knowledge refers to an "individual's mental models consisting of mental maps, beliefs, paradigms, and viewpoints."...explicit knowledge refers to "articulated, codified and communicated...in symbolic form and/or natural language" (Alavi and Leidner, 2001). In just these two dimensions we find ourselves asking the following questions:

- Can viewpoints and beliefs not also be made explicit?
- Can mental maps not also incorporate symbols?
- Can explicit knowledge be specialized knowledge?
- Can tacit knowledge be generalized knowledge?

The answer to each is "yes". First, viewpoints can be made explicit; we see it in the form of newspaper "opinion editorials" (Katz, 1959). Second, human mental thought has been described as being symbolic (Zilhao, 2011). Third, Einstein's theory of special relativity, while specialized, was widely printed (De sitter, 1917). Finally, tacit knowledge need not be specialized, or only locally transferrable (Gertler, 2003; Friedman, 1987). Notwithstanding the inconsistencies, is there not a way to measure the evolvement of information within an information system? Physics points us towards a “science of information” (Brukner and Zeilinger, 2005). Biological information systems points us towards an evolutionary perspective of data (Kay, 1998). Herein we introduce our theory of dataphoric space.
Theoretical Framework of Dataphoric Space

Dataphoric space is an interdisciplinary framework that explores the utility of an integrated theory inspired by population ecology, information theory and space-time fabrics. A space-time fabric is the underlying foundation of dataphoric space. While this fabric is physical in nature, it has a measurable information geometry (Minkowski, 1952; Shen, 2006). This geometry can be described via descriptive features, statistical features, entropic features and model-based features (Zhongguo, Hongqi, Ali, and Yile, 2017). Furthermore, the intersection of ecology and information theory is not without precedence (Margalef, 1957; Ehrlich and Holm, 1962). It is through this interdisciplinary perspective that we developed our integrated theory of dataphoric space (See Figure 1. Dataphoric Space).

(Figure 1. Dataphoric Space)

Cellular Data Forms

Cellular data forms are singular units of information. For example, if we have a singular letter, the letter “A”. As an individual “cellular” data form, it has a small entropy value when compared to a string of letters “ABCD”. Furthermore, when a string of letters is constructed into a word, that word may have the same entropic value between the two different string values. For example, the string “ABC” and the string “CAB” would both have the same entropy value. In addition, the sentences on this page would have a higher entropy value versus the words themselves. Consequently, the entropic value of a singular page within a book is higher than the individual datum themselves within the sentences.

Histological Data Forms

Histological data forms are a combined aggregation of cellular data forms. As cellular data gains more mass within a dataphora it will eventually combine into more complex data forms. These data forms can best be described as small to large sentences. Complex word combinations that form the beginnings of an idea. Histological data forms in and of themselves are not useful as they have only enough content to describe the most basic of concepts. Many multiple histological data forms are required to describe more complex concepts. Continuing with our example of document dataphores, a dataphore is created to the degree where a common paragraph of information is created from many multiple histological data forms.

Dataphora

A dataphora is a biome of information. A dataphora is a region of dataphoric space that is massive in content (a kind of gravitational data density). Entropically we would describe it as being high (i.e., containing more information complexity). Dataphores reside within a dataphora - this is equivalent to a species, or flora within a biome. For example, laptops, servers, cell phones, documents, spreadsheets, databases – and even biological entities (including humans) – are all dataphores operating within a dataphora. Additionally, the concept of a dataphora scales depending on the research question at hand. For example, a hospital, a country, the entirety of the planet or even our galactic local group could be viewed as a dataphora.

Dataphores

Dataphores are composed of many histological data forms (HDFs). We elected to use “histological” as a means to describe tissue, not quite an organ of data, but a type of biologically inspired “data tissue”. Data in turn is composed of cellular data forms (CDFs) called datum. Again, biologically inspired in that “cellular” is one of our lowest levels. CDFs could be broken down into data units called "dats". We include dats, as a level below that of a datum will be useful for future research.
Entropic Values

The famous physicist Leonard Susskind described data contained within a tub of water in the following way, "there is an enormous amount of hidden information...hidden information that you can't ordinarily see...that hidden information is called entropy..." (Susskind, 2007). Yet, we can use entropy in a general sense to describe data-centric phenomenon. For example, a textbook will have a higher entropic value versus a string of redundant letters (Shannon, 1948). In this way, we can use entropy to contextualize data entities within dataphoric space. Our data shows that higher values of entropy describe higher levels of information complexity (i.e. more data equates to more data uniformity).

Research Method

Our research is comprised of two studies: a data study and a semantic study (See Figure 2. Research Instrumentation). Our data study defines the ontological direction of this research and establishes a baseline for human sensemaking of document dataphores as a function of data entropy and data geometry. The semantic study drops down a level by focusing on human sensemaking as a function of semantic compression as well as the relational distances of content within a document dataphore.

The data study portion of our research is grouped into two constructs: data entropy and data geometry. Herein we detail the entropic portion of our ongoing research effort.

Data Collection

We randomly collected 72 content samples using Wikipedia. These samples ranged in size from tiny single character units of information to very large articles. Using this information we were then able to leverage Shannon's calculation for entropy to calculate the entropic value based on the information content being supplied (See Table 1. Sample Entropy Data Set). Lower entropic values are statistically speaking, predictable. Lower values of entropy are less noisy, higher entropic values are less predictable, they are more noisy. In a way it is a measure of information complexity and relative uniformity.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Data</th>
<th>Data Length</th>
<th>Entropic Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiny String</td>
<td>a</td>
<td>1</td>
<td>1.0000</td>
</tr>
<tr>
<td>Integer</td>
<td>1</td>
<td>1</td>
<td>1.0000</td>
</tr>
<tr>
<td>Two Words</td>
<td>hello world</td>
<td>11</td>
<td>2.6464</td>
</tr>
<tr>
<td>Big Word</td>
<td>accoutrements</td>
<td>13</td>
<td>3.2389</td>
</tr>
<tr>
<td>Three Words</td>
<td>formula for entr</td>
<td>19</td>
<td>3.4104</td>
</tr>
<tr>
<td>Small Article</td>
<td>In statistics, max</td>
<td>1394</td>
<td>4.2169</td>
</tr>
<tr>
<td>Large Article</td>
<td>After the release</td>
<td>7543</td>
<td>4.5075</td>
</tr>
<tr>
<td>Large Article</td>
<td>Lacaillle gave Bay</td>
<td>6072</td>
<td>4.6793</td>
</tr>
</tbody>
</table>

After completing the entropy calculations for all data samples, we then analyzed the entropic data values using an unsupervised k-means clustering algorithm. This is a common method for doing exploratory data analysis. Clusters produced via k-means are divided by ‘n’ number of observations in which the observations are nearest a given number of centroids (i.e., many statistical means).

We used the data mining tool Orange3, a component-based data mining tool useful in employing research grade supervised and unsupervised machine learning techniques on medium to small data sets (Altalhi, Luna, Vallejo, and Ventura, 2017). Using Orange3 we experimented with the number of centroids supplied to our clustering model. In short, we are looking for the most equitable distribution of clusters (i.e., balance across the means). Using data length and calculated entropic values we discovered that twelve distinct clusters was the most descriptive (See Figure 3. Clustering via k-means).
Data Observations
The lower two clusters correspond to cellular data forms and the middle two clusters correspond to histological data forms. There is a transitory layer that appears between histological data forms and dataphores. This was an unexpected find within the data. This transition occurs where entropic values are between 3.6 and 4.2 (i.e., the red cluster). We believe that histological data forms gather “data momentum”. This momentum leads to a greater influx of information, a type of gravitational data pull (Walker and Alrehamy, 2015). These transitional data forms are a bridge of sorts, akin to moving a simple idea written on a napkin over to a more formalized document. Additionally, cellular data forms are generally categorized as having entropic values between 0.0 and 2.2; histological data forms between 2.2 and 3.6.

Dataphores are categorized as having entropic values above 4.2; the dataphores in this case represent small to large Wikipedia documents. As a dataphore grows in content, the increased complexity as a function of entropy tapers off drastically. Therefore, the clustering at this point was based entirely on data length. Per the entropy calculation, we surmise that as information content rises beyond a few hundred characters, it has less of an effect on entropy values due to the logarithmic nature of the Shannon’s calculation of entropy.

Limitations
First, we are working with a small sample size. Second, when using entropic values, mathematical equations and complex string sequences can have abnormally high entropic values. For example, an alphanumeric string sequence of “abcdefghijklmnopqrstuvwxuyz0123456789” has an entropic value of 5.16; this is surprisingly high given that the data length is only 37 characters long. Of course, this does not elevate this data string to the level of a being a dataphore even though the entropic value is very high. Therefore, the usage of Shannon’s entropic value, combined with data length must be used for what we call “common content scenarios”. Consequently, a researcher must keep an eye towards cleaning up any abstract knowledge constructs that create outliers within the data.

Conclusion & Implications
This research introduces the theoretical concepts of dataphoric space. We have determined that entropy is usable in categorizing data elements within the dataphoric research ontology. This research could be extended to other dataphores (e.g., videos, databases, applications, artificial intelligences and even humans). However, we tend to lose categorical fidelity when using entropic values at larger scales. Extending the framework to calculate entropic values at a dataphora-level requires us to expand on Shannon’s entropy calculations as we would like to magnify our ability to distinguish entropically between macro, mezzo and micro structures within dataphoric space. Ultimately, this research presents the concept of biological speciation of data within an information system. Historical terminology can have an effect on our cognitive thought process (Maslow, 1969). There is a grain of truth in the Sapir-Whorf hypothesis, the terminology we learn as children to describe various phenomenon around us does have an impact on how we compartmentalize information (Cibelli et al., 2016). Another path of inquiry becomes available if we allow ourselves to view information systems as biologically styled entities. Using this ontology, information systems researchers will be able to create, observe and analyze the evolution of dataphores within a dataphora across multiple domains of scientific inquiry (e.g., sociological, anthropological, biological, physical, medical...). As we move forward, we expect that our dataphoric terminology will expand over time to encompass more constructs. Ultimately, we see dataphoric space as a stimulating development for information systems research.
REFERENCES


