Taxonomy in Biology and Visualization

Eamonn Maguire

Oxford University Department of Computer Science

Abstract—Taxonomies are core to our everyday working. We constantly categorize entities to aid our understanding and navigation of the world around us. In the most basic of cases, we may categorize animals into dangerous or non-dangerous, pets or pests and people into friends and strangers. Since time began, our ancestors have used taxonomy to stay away from things carrying the potential to cause harm, and it is shown that we have the ability to classify complex entities even as children through evolved parts of the human brain. This paper explores the origin of taxonomy and how it carries immense importance in our world, its importance in biology and how it is used in visualization also.

Index Terms—Taxonomy, History, Biology, Visualization.

1 What is taxonomy?

> **Taxonomy**: from Ancient Greek, *taxis* **"arrangement"** and *nomia* **"method"** [20]

A *general* definition for **taxonomy** from Simpson [31] is:

> "A field of science (and major component of systematics) that encompasses description, identification, nomenclature (naming), and classification."

There has always been a need to identify, name and classify the entities around us. Through identifying and naming animals, insects, arachnids, plants, shapes or objects, we can adequately communicate with others about the environment around us. The classification of the entities found within our environment, arranged in a hierarchical format are referred to as taxonomies. Taxonomies construct *our* view of the world and our interactions within it. We constantly identify things around us, even though we may not know the scientific name for them. We also classify these things: things with fur, hair, nice shoes or nice clothes. We stay close to things which are positively reinforced and avoid things with negative connotations [35]. For instance, even though one doesn't personally know every homeless person in Oxford, most people can very quickly spot a homeless person given general characteristics such a clothing, behaviour or facial complexion. This view we concoct is known as the umwelt, the German word for "environment" or "the world around" [37, 35, 22]. Within each of our own *umwelten* (plural) we have different views, and often different ways of classifying things (although a lot of things will be classified in the same way).

Over the years man has tried to name and classify the environment around using differing approaches. Of particular interest was classification of the organisms living around us and this is where taxonomy, in the documented sense, has its origins. Through naming and classifying the organisms around us we could more clearly communicate which things were dangerous and which were not. We could also see which organisms were the same type, which things all had fur, four legs, two eyes and so forth. In this chapter we describe the many approaches taken to devise such a taxonomy by luminaries such as Aristotle and Carl Linnaeus, often referred to as the fathers of science and taxonomy respectively. We also assess the use of taxonomies within biology, not just for classifying organisms but for classification of chemicals too for instance. Additionally, we assess the role of taxonomy in the field of visualization where it is playing an increasingly important role in classifying types of visualizations to make it easier for researchers and users alike to search through the cornucopia of resources currently available but not easily accessible.

• *Eamonn Maguire is with Oxford University, E-mail: eamonn.maguire@st-annes.ox.ac.uk.*

Manuscript received 31 March 2011; accepted 1 August 2011; posted online 23 October 2011; mailed on 14 October 2011.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

2 Taxonomy in Biology

> "Whats the use of their having names," the Gnat said, "if they won't answer to them?"
>
> "No use to *them*," said Alice; "but it's useful to the people that name them, I suppose. If not, why do they have names at all?"
>
> "I can't say," the Gnat replied.
>
> **- Lewis Carroll**
> *Through the Looking-Glass*

2.1 The history of taxonomies. It all started with biology

Since the origins of taxonomy lie within the field of biology, it is worthwhile to look at the history of how taxonomy evolved within this field and its major contributors. The timeline for the history of biological taxonomies can be split into three key periods revolving around the man regarded to be the "father of taxonomy", Swedish botanist Carl Linnaeus (1707-1778). These periods are pre-Linnean, Linnean and post-Linnean.

2.1.1 Pre-Linnean *(3000BC - 1707)*

The concept of taxonomy did not originate with the Roman and Greek cultures as is thought by the majority of people. In fact the earliest traces of taxonomy are from the Eastern world rather than the West [23].

Shen Nung, Emperor of China (3000 BC)
The earliest form of document resembling a taxonomy dates back to 3000 BC in a pharmacopeoia named *Divine Husbandman's Materia Medica* written by Shen Nung, Emperor of China. It consisted of 365 plant, mineral and animal derived medicines[23]. Further to this, in Egypt around 1500 BC medicinal plants were illustrated in wall paintings [23]. Additionally, one of the oldest papyrus rolls, named Ebers Papyrus, classified plants by their ability to treat various diseases of that age[23].

Aristotle (384 BC - 322 BC)
Between 384 BC and 79 AD, the Greeks and Romans began to take an interest in taxonomy. Aristotle, commonly known as the "father of science"[23] was the first in the Western society to classify all living things in *Historia Animalium*. His classification scheme was hierarchical and grouped organisms according to whether or not they had similar physiological, behavioural and morphological

characteristcs, for instance animals with and without blood, animals who live on water and those living on land. His classification system had no concept of evolution and the species within the hierarchies had no known genetic relationship since this wasn't known, they just looked/behaved in a similar way. Aristotle also introduced two important concepts of modern taxonomy, those being: binomial (2 name) definition (e.g. genus species names such as *Homo sapiens* later devised by Linnaeus); and classifying organisms by type[23].

Theophrastus (370 BC - 285 BC)
Aristotle was succeeded by his student Theophrastus who extended his tutors work through the classification of all 480 known plants in *De Historia Plantarum*. He subsequently became known as the "father of botany". His classification scheme was based upon characteristics such as growth habit and fruit/seed form, introducing categories such as trees, shrubs and herbs [23]. Many of the genera names Theophrastus developed were accepted by Carl Linnaeus [23].

Plinius (23-79 AD)
After Theophrastus, 318 years passed before Plinius arrived who wrote *Naturalis Historia*. His main contribution to taxonomy was the introduction of Latin names, a contribution for which he became known as the father of botanical Latin [23].

Dioscorides (40-90 AD)
Shortly after the arrival of Plinius, there was Dioscorides who gathered information about medicinal plants and wrote *De Materia Medica*, a publication containing approximately 600 species classified by their medicinal properties[23].

Caesalpino (1519-1603)
Until the development of optic lenses towards the end of the 16th century [23], the works of the Greek and Romans remained intact and there was little in the way of further classification mechanisms developed. Microscopy enabled discovery of a greater number of species as the level of granularity at which scientists looked at the specimens became more and more refined. Higher granularity meant that simply classifying things based on characteristics such as petal colour, leaf type and so forth was no longer enough. Microscopes allowed scientists to investigate many more intricate characteristics of organisms which led to an immediate increase in species count, since some things once thought to be the same were now different based on these newly visible characteristics. The first of this new breed of taxonomists aided by this new technology was an Italian named Caesalpino whose work named *De Plantis* contained 1500 species, considerably more than Theophrastus[23].

Bauhin brothers (1541-1631; 1560-1624)
The Bauhin brothers, from Switzerland wrote *Pinax Theatri Botanici*, a revolutionary volume containing 6000 species as well as their synonyms, a novel feature required to bring together alternate namings of the same species (e.g. Homo Sapiens and Human in today's day and age) [23]. Moreover, through grouping species by their genus and species, the Bauhin brothers were the first to use the binomial nomenclature first introduced by Aristotle.

John Ray (1627-1705)
The Bauhin brothers were followed by John Ray who in 1682 published *Methodus Plantarum Nova* containing around 18,000 plant species classified taking into consideration many more characteristics of the organisms. He made great strides in his pioneering work concerning entomological taxonomy[23].

Joseph Pitton de Tournefort (1657-1708)
The final taxonomist to come along before Carl Linnaeus was Frenchman Joseph Pitton de Tournefort) who in 1700 published *Institutiones Rei Herbariae* containing a taxonomy of 9000 species listed in 698 genera classified by their floral characteristics (most of which were accepted by Linnaeus and still in use today) [23]. It became the reference taxonomy for that botanists up until Carl Linnaeus' publication named *Systema Naturae* in 1735.

2.1.2 Linnean *(1707 - 1778)*

Carl Linnaeus was a Swedish physician who placed botany as a focal part of his study. When Linnaeus was born in 1707, there was no shortage of botanical classification systems in use. However, there was no one size fits all naming convention or classification method developed and the result was no consistency where the same plants had numerous names. The Bauhin brothers applied a pragmatic yet short term approach to resolve this problem through the introduction of synonyms to species descriptions, but as new species were continually being discovered as world exploration reached a new level, the situation was quickly becoming unmanageable. In an attempt to quell the potential name epidemic, Linnaeus published *Critica botanica* in 1735 which presented guidelines to be followed for creation of generic names, extending Aristotle's introduction of binomial definition into the naming scheme of the world's flora and fauna [23]. This two part name was constructed from the genus (always capitalized) and the species name (always lower case), e.g *Homo sapiens*. Further to this, Linnaeus realized the need to not only harmonize the names of species, but to also control the way they are described. Linnaeus introduced rules for how to construct species description and the terminology to use in two publications, *Fundamenta botanica* in 1736 and *Philosophia botanica* in 1751 [23].

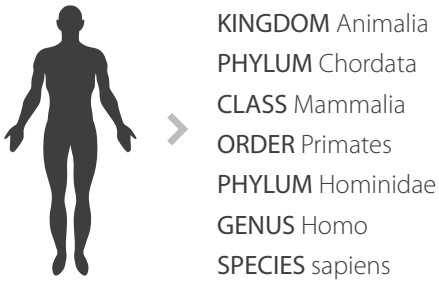

Fig. 1. Linnaean Classification of *Homo sapiens*.

In developing a common classification method, Linneus wrote his most important work *Systema Naturae or The System of Nature*. Published in 1735, it provided an overall classification framework for all plants and animals from kingdom to species level (**Kingdom - Phylum - Class - Order - Phylum - Genus - Species** *see figure 1 for an example classification*) [33, 23]. Over the course of 10 editions, the last of which was published in 1758 and whose full title was *Systema naturae, sive regna tria naturae systematice proposita per classes, ordines, genera, and species*, Linnaeus fixed errors such as the positioning of whales as fish and he was the first to put humans amongst primates and classify us using the binomen *Homo sapiens*. Further to *Systema Naturae*, Linnaeus' plant classification, influenced by Cesalpino, was described in two publications named *The Genera of Plants* and *The Species of Plants*. His classification method built upon Tournefort's method of classification, however Linnaeus extended this method considering that plant's had a sexuality based on the presence of stamens and pistils [23].

In his quest for perfection, Linnaeus became the "father of taxonomy". His constant refinement of his methodology based on emerging species and developments in science enabled the pervasive use of his classification methods. Moreover, his foresight in being able to not only solve many of the problems of that time in creating common classifications, but also to introduce standard naming conventions and species descriptions was critical in making the science of taxonomy what it is today.

2.1.3 Post-Linnean

Many of the taxonomies we use today are still based on Linnaeus' work back in the 18th Century. However, back in those times, not everyone agreed with Linnaeus' classification approach, the French in particular. Disagreements were in general justified. Linnaeus's classification approach was very subjective and categorised things based on observed phenomena, such as whether or not an animal had fur, hair, 4 legs and so forth. Four French scientists, from 1707 - 1829 including George-Luise Leclerc de Buffon (1707-1788), Michel Adanson (1727-1806), Antoine Laurent de Jussieu (1748-1836) and Jean-Baptiste de Lamarck (1744-1829) challenged the ideas of Linnaeus and bettered the taxonomic field as a whole through contributions including new theories about evolutionary traits for example.

George-Luise Leclerc de Buffon (1707-1788)
George-Luise Leclerc de Buffon's criticisms were based on the fact that Linnaeus was imposing order on a chaotic natural world. He developed theories about how species develop, varieties within species and inherited characteristics in species [23]. This work provided the platform on which evolutionary biology was based.

Antoine Laurent de Jussieu (1748-1836)
Antoine Laurent de Jussieu bettered Linaeus' classification of plants with *Genera Plantarum* in 1789. He launched a natural classification system based on many plant characteristics which are now used in modern classification systems [23]. He introduced the classifications of *acotyledons*, *monocotyledons* and *dicotyledons* and added the *"family"* rank in between *"genus"* and *"class"* [23].

Jean-Baptiste de Lamarck (1744-1829)
Jean-Baptiste de Lamarck developed an evolutionary theory which also included the inheritance of characteristics between species [23].

Darwin era (1809-1882) and evolutionary theory
Although the notion of evolution was presented before Charles Darwins time by Lamarck and de Buffon for instance, it was not fully explored until Darwin and Alfred Russel Wallace (1823-1913) launched the *evolutionary theory* in 1858. Following this, two German biologists, Ernst Haeckel (1824-1919) and August Wilhelm Eichler (1839-1878) composed ideas to incorporate evolution into taxonomies. It was Haeckel who eventually established the term *"phylogeny"* leading to what we now know as phylogenetic trees which visually group similar things together based on the presence/absence of characteristics between organisms or more recently their genetic distance.

Willi Hennig (1913-1976) and cladistics
German biologist Willi Hennig (1913-1976) founded a new classification method called **cladistics (or phylogenetic nomenclature)** in 1966 which was intended to be a more objective method for classification organisms. Cladistics stated that only similarities grouping species, called synapomorphies should be used in classification and taxa should include all descendants from one single ancestor (monophyly). An example cladogram is shown in *figure 2*.

The problem with phylogenetic trees is that they typically yield a high number of possible trees for any given number of taxa and subsequent characteristics, be these genetic or phenetic: 3 taxa yield 3 trees; 4 taxa yield 15 trees; 7 taxa yield 10,000 trees; and with 10 taxa, there is a possible 34 million trees [32]. In an attempt to resolve this, cladistics introduced *parsimony*, where the best phylogenetic tree is one which requires the least number of evolutionary changes. An example is worked through in *figure 3*.

The value of cladistics wasn't realised by everyone when it was initially proposed. When Willi Hennig suggested clasdistics as an approach, scientists were more interested in phenetics, which dominated for years after. Phenetics involved grouping organisms by many different characteristics and was perceived to be much more subjective than classic Linnaean classifications. It wasn't until when PCR (polymerase chain reaction) and DNA sequencing started to come on the scene in the 70s, that the cladist approach started to make more sense and picked up momentum. Additionally, calculating 34 million possible trees using computers in the 70's wasn't very feasible unless one had access to large university supercomputers, and this is only a small example with 10 taxa! Building up a complete tree of life or even clade takes a huge amount of computation which until relatively has not been possible.

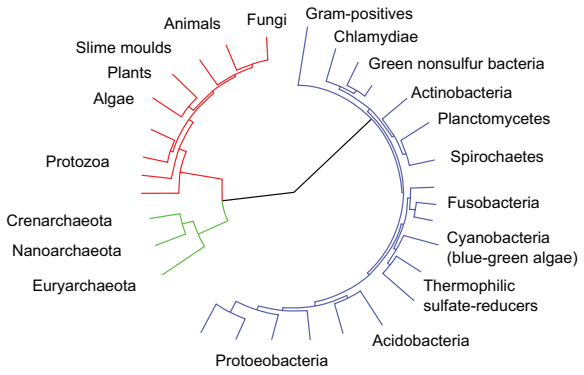

Fig. 2. Tree of life as a cladogram. Source: Wikipedia http://en.wikipedia.org/wiki/File:CollapsedtreeLabels-simplified.svg

Moving toward cladistics & PhyloCode (1970s - present)
As a result of this proliferation of sequences and subsequent computational analysis, we'll find that lots of organisms don't belong in their Linnaean classification any longer. With Linnaean classifications, new discoveries can require renaming of classes, orders or kingdoms (or combinations of all three). Over the last 30 years, there has been continuing refactoring of the classification tree to make the branches monophyletic (one common ancestor). This will continue to happen as we continue to receive more sequence data. Conversely, the cladist approach was developed with the assumption that the shape of the tree can change, making it inherently less difficult to add/remove/change classifications as time goes on.

Even though cladistics is in use today, it is again not without it's critics. For instance, when comparing species based on genetic traits and genetic distance where our matrix for parsimony is now based on the probability that adenine can change to thymine, guanine can change to cytosine and vice versa, we fail to take into consideration well know concepts such as lateral gene transfer or events such as those where bacteria cells can have viral DNA. As a result (as unlikely it might be), our classification can be wrong and possibly place a bacteria in the same clade as a virus, simply because they share a large proportion of the same DNA. There is also a criticism in the use of parsimony. The assumption that the best tree is the one with the fewest number of evolutionary steps is not necessarily true since evolution doesn't always do things in a straightforward manner.

Even so, there are still movements towards cladistics and away from Linnaean nomenclature and classification. Borne within these movements, the PhyloCode project [3] started in 1998. The concept of PhyloCode is that species and clades should have names, but all ranks above species are excluded from the nomenclature [3, 23]. The project is regarded as controversial within the bio taxonomic community, but represents the latest suggested change in how classifications could be created.

2.2 Other taxonomies within biology

Taxonomies or ontologies (which are more formal than basic taxonomies with rules, restrictions, etc.) are used in many parts of biology in order to structure domain knowledge in order to aid scientists in consistently identifying chemicals or genes for instance:

- biochemistry: naming and describing chemicals and small molecules, and classifying these (see figure 4). See the chemical entities of biological interest ontology *ChEBI* [1]; and

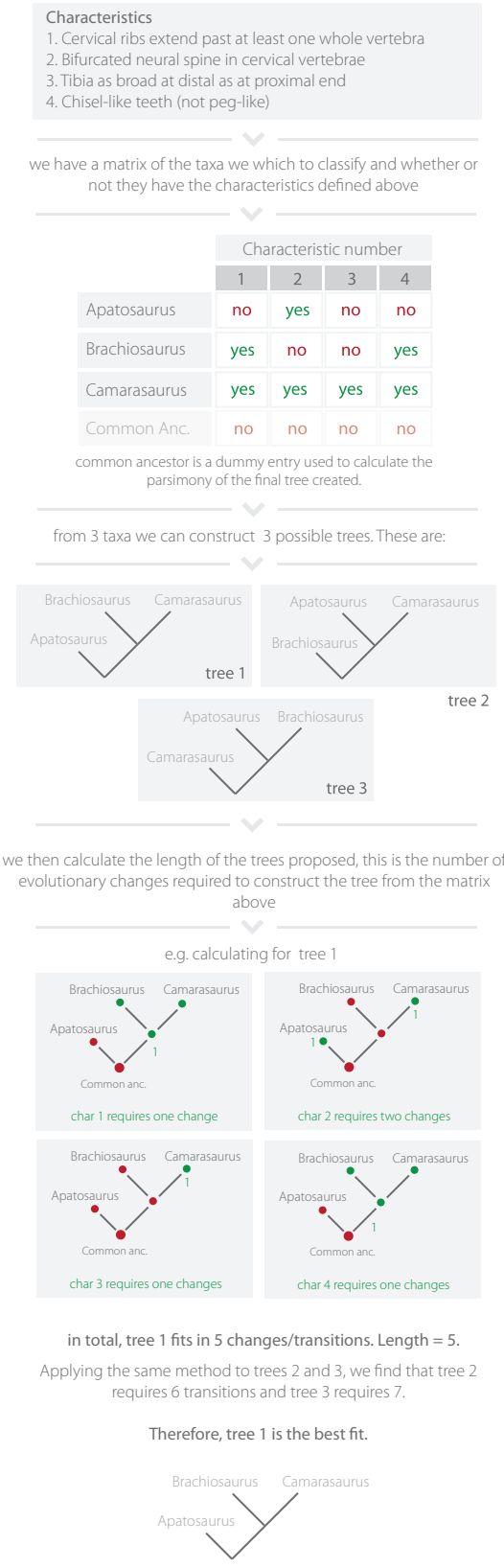

Fig. 3. Characteristic matrix is constructed given 3 taxa. 3 taxa yields 3 possible trees and the best tree is selected based on the one that requires the fewest number of changes to be constructed. Diagram constructed based on content from Taylor [32].

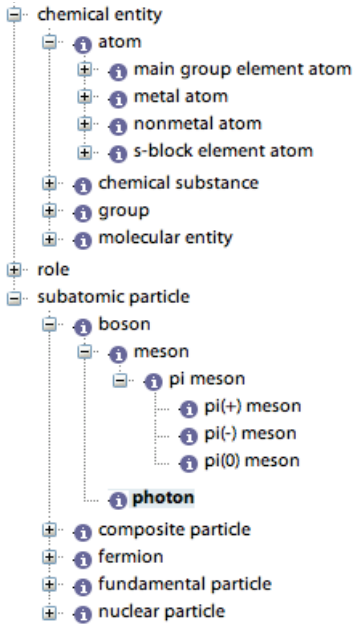

Fig. 4. High level classifications in the ChEBI ontology [17, 1], as viewed in BioPortal

- genes: describing the biological processes, cell location and molecular functions of genes. See the gene ontology in BioPortal *GO* [15, 2].

Consistently describing organisms, chemicals, genes, biological processes and so forth is very important in allowing for accurate dissemination of knowledge.

2.3 Summary of biological taxonomies

Every taxonomy created will come under scrutiny as a result of the nature of the task at hand. We are asking scientists to map continuous phenomena in terms of evolution and speciation to discrete space which is inevitably going to be challenging and always artificial [21]. Moreover, as more and more DNA sequencing continues to happen, there will be many shifts in the tree of life as there was when microscopes first became available back in the 16th century.

Even though criticisms of taxonomies (and ontologies) exist and are often far from complete or accurate, they are also hugely important in our understanding of biodiversity and the world around us. For instance, in human metagenome studies, we can identify microbes living in and on our bodies and the effects of such populations on health[9]. In microbial ecology, we can discover the species prevalent in particular habitats, for example, which bacteria are present in areas where an oil spill has occurred [19]. Having well defined taxonomies and information about the species, studies like those described can happen. Without such taxonomies, we would have no idea what exists in the environment around us.

Extending on this point, as we continue to sequence and study more and more of our ecology and personal microbiomes via initiatives such as the human and earth microbiome projects [26], challenges are being encountered when defining when exactly a new species has been found. There is still a lot of uncertainty over how to define a new species and how much natural in-species variation is allowed [21, 24]. There needs to be more work carried out to make systematic decisions about what is a new species and what is not. From 1993 - 2005, the number of mammalian species alone has grown from 4,659 to 5,418 [25]. Given that the microbial populations are magnitudes in size bigger, we are set for a taxonomic explosion unless a more systematic approach is put in place.

3 TAXONOMIES IN VISUALIZATION

The idea of a visualization taxonomy first appeared in 1992 and was introduced by Brodlie from the University of Leeds [8]. Since then there have been numerous taxonomies developed to try and capture what happens within the visualization field. We typically wish to categorize the following areas of visualization:

- 1. the type of visualization, e.g. the data being represented and the way we've represented it;
- 2. how we created the visualization (processes), so the assumptions that were made when abstracting the information from the initial data source and when creating the visual mappings; and
- 3. the organisation of visual taxonomies and subsequent visual metaphors to represent real world concepts (semiotics).

1 and 2 are very much related and are mechanisms for categorising visualization techniques in general. 3 on the other hand diverges from the normal idea of classifying things within visualization such as data or chart types and instead focuses on how one may leverage off of the internal taxonomies we carry with us to create icons/signs/glyphs which map to real life phenomena, e.g. a cross on a map corresponds to a church or a red H maps to a hospital.

The end result is that a visualization taxonomy should be able to:

- **help users** in navigating the huge visualization tool space. By adequately categorising visualizations users could more easily find the visualization that is right for them; and
- **help researchers** in finding out about similar work in their field or work that can add value to their research.

3.1 Taxonomies for visualization techniques

Two ways currently explored for the creation of visualization taxonomies are to:

- classify by the type of the data being visualized, e.g. discrete or continuous data. This is the most common way to classify visualizations; and
- classify by the processes/algorithms used to process and render the data, a relatively new method of classification.

3.1.1 Focusing on visualization data type

Most of the taxonomies created in the past have focused on data type. The categories defined have historically made the split on discrete vs continuous or scientific vs non-scientific data, indicated by factors such as; application area; whether the data is physically based (*scientific visualization*) or abstract (*information visualization*); and if spatial information is given (*scientific visualization*) or not (*information visualization*) [34]. The problem with such a taxonomic hierarchy is that there are many examples which cross both classifications, so classifying something simply because it has to go somewhere would be inherently wrong. Below we document some of the major developments in visualization taxomonies and look at how they are structured.

In 1996, Shneiderman described a taxonomy for visualization which was data centric but also included a taxonomy on how to navigate through the visualization space based on a "visual information seeking mantra" of *"overview first, zoom and filter, then details on demand"*[29]. The taxonomy is arranged as follows [29]:

- Data types:
  - 1-dimensional e.g. textual documents & program source code;
  - 2-dimensional e.g. geographic maps & floor plans;
  - 3-dimension e.g. molecules & the human body;
  - multi-dimensional e.g. items with n attributes which become points in a n-dimensional space;
  - temporal e.g. project management & historical representations of data;
  - tree e.g. file hierarchies;
  - network e.g. protein-protein interaction networks & social networks;
- Interactions,:
  - Overview;
  - Zoom;
  - Filter;
  - Details-on-demand;
  - Relate;
  - History;
  - Extract;

The visualization taxonomy from Shneiderman has been applied by his students to create an online resource called the *On-line Library of Information Visualization Environments (OLIVE)* [4] which categorizes projects, products and publications by the data types defined in the taxonomy.

In 1997, Card and Mackinlay devised another data-oriented taxonomy [11] which was then expanded on in 1999 in a collaboration with Shneiderman [10]. The overall contribution of this work was to split visualization into a number of categories. These were:

- Scientific Visualization;
- GIS;
- Multi-dimensional Plots;
- Multi-dimensional Tables;
- Information Landscapes and Spaces;
- Node and Link;
- Trees;
- Text transforms;

In 1998, Ed Chi proposed a taxonomy focusing on not only the data, but also the processes these data goes through to create a visualization [14]. This taxonomy is called the Information Visualization Data State Reference Model (or Data State Model). A graphical depiction of the taxonomy is shown in *figure 5*. The concept is to classify visualizations on the type of data they work on (temporal, network, multi-dimension etc) and the transformations such data goes through to create the final visualization including data abstraction, visual abstraction and visual mappings.

The generic nature of Chi's taxonomy makes it able to classify many visualization techniques and numerous examples are documented in Chi's follow up publication in 2000[13]. The concepts described in Chi's paper have also been applied in creating more domain specific taxonomies as exemplified by Daasi *et al* [16].

Furthermore, have been additional visualization taxonomies suggested in some non-scientific publications which categorise visualizations by the type of operation they support For instance, Dan Roam presented the Visual Thinking Codex (*see figure 6*) in his book named *The Back of the Napkin* [27].

In this figure Roam organises visualization types by what they are trying to divulge to the reader of the visualization: who/what, how much, where, when, how and why? Additionally, there are subcategories on if the visualization is simple vs elaborate in terms of detail or qualitative vs quantitave for example.

Andrew Abela [5] also presented a technique for taxonomising visualizations. This is shown in *figure 7*. The high level (comparison,

Fig. 5. Taxonomy for visualization, derived from Chi [14, 13]

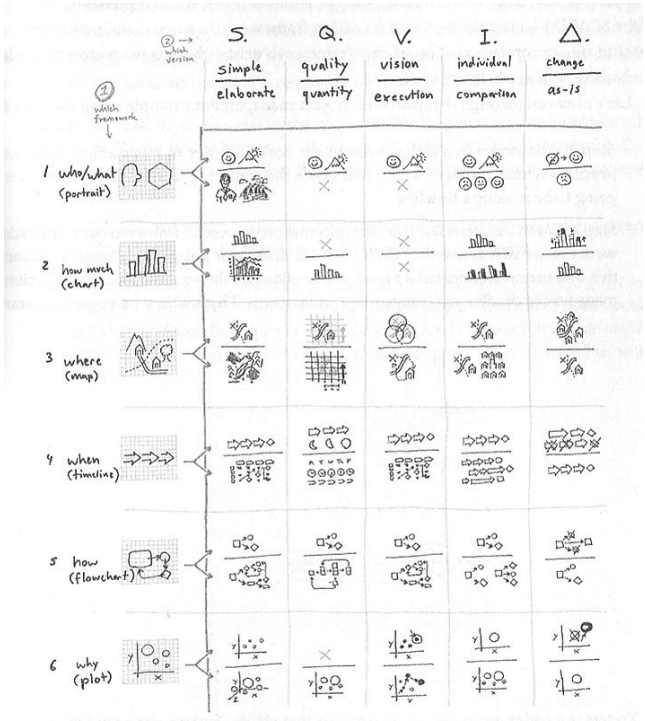

Fig. 6. The Visual thinking codex from pg. 141 of Dan Roam's *The Back of the Napkin*.[27]

relationship, distribution and composition) of the taxonomy classifies based on what the visualization is intended to show. There are sub-classifications for each of these taking into consideration the number of variables and whether or not data is static or dynamic. Additionally, classifications identified by Abela have been used in developing Chart Chooser (http://www.juiceanalytics.com/chart-chooser/) created by JuiceAnalytics, however the taxonomy used here is a simplified version of Abela's and doesn't cover the level of detail required to sufficiently categorise all visualizations.

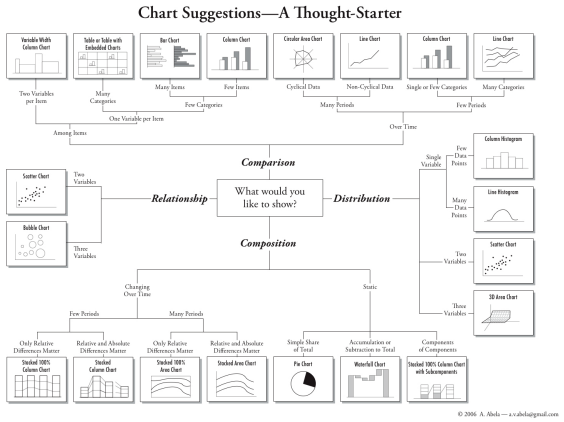

Fig. 7. Taxonomy created by Andrew Abela [5]

3.1.2 Focusing on data processing algorithm for visualization

In 2004, Tory and Mller published a comprehensive taxonomy [34] extending work already carried out by Card and Mackinlay [11, 10], Shneiderman [29] and Chi [14, 13] to create a comprehensive taxonomy for visualization. The taxonomy classifies visualization algorithms (as opposed to the data in almost all other visualization taxonomies) and the assumptions these algorithms make about the data to be visualized [34]. The taxonomy is organized as follows [34].

- Discrete/Continuous algorithms;
  - Discrete algorithms assume that data *can not be* interpolated. Within this category, there is a further classification based on whether or not the model is structural or value based where further classifications are made on the dimensionality. This is illustrated in *figure 9*;

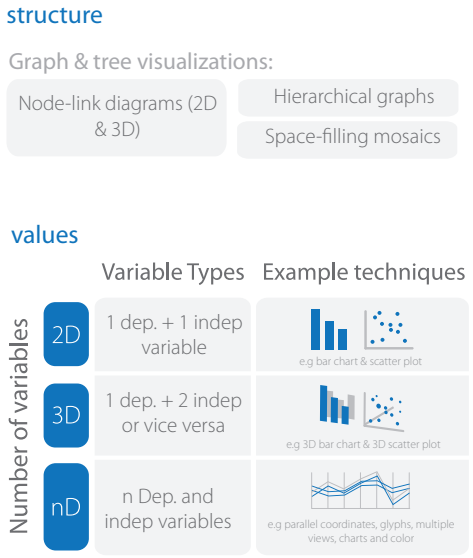

Fig. 8. Low-level taxonomy of discrete models from [34]

  - Continuous algorithms assume that data *can be* interpolated. Within the category there is further classification

based on the data structure (scalar, vector, tensor or multi-variate and the number of independent variables.);

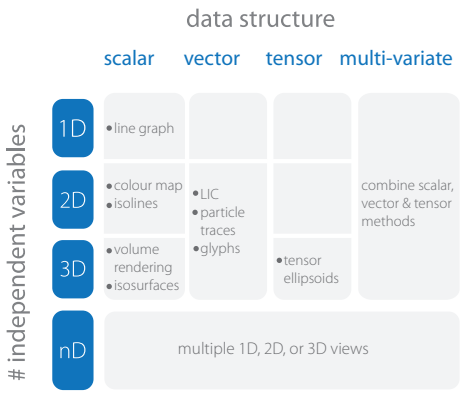

Fig. 9. Low-level taxonomy of continuous models from [34]

- Display Constraints. Discrete and continuous algorithms are further classified by whether or not the display constraints are given, constrained or chosen.;
  - Given, e.g. air traffic positions (discrete), or images (continuous);
  - Constrained e.g. arrangement of ordinal/numeric values (discrete), or 2D geographic maps (continuous);
  - Chosen e.g. file structures (discrete) or regression (continuous);
- Visualization Tasks. For each type of visualization, there will be different types of tasks that can be performed, something that wasn't detailed in the taxonomy created by Shneiderman [29] *(see figure 10)*;
  - When spatialization is given: user should be able to specify regions of interest, extract information and/or examine information in greater detail;
  - Discrete structural models allow pattern analysis: user can identify outliers and clusters of data points as well as filter data;
  - Continuous models (and discrete value models holding ordinal data) allow users to see numerical trends;

There is much novelty in the algorithm compared with other techniques focusing on just the data type. However, in my opinion there is still fuzziness in how the classification is made at the top level as demonstrated in *table 1* of the paper [34] where *display attributes constrained on spatialization* yield duplicated examples in two separate classifications.

3.2 Taxonomies and semiotics

All seeing humans, animals, insects and so forth have a visual taxonomy, things are seen and meaning is constructed from previous inferred or known knowledge. For instance, we see a red sign on a road and immediately take notice since we know that red generally means 'warning'. Similarly, if we see an insect with bright colours (red in particular), it is natures way of telling us to watch out, that insect is most probably poisonous (or pretending to be). Interestingly, noone has taught us that red means danger, it is known. Red is in our internal taxonomy, there since we were born to tell us that red things are usually dangerous. The relations between the visual and their meaning has been investigated many times over by authors such as Jakob Von Uexkll [35], Jacques Bertin [7, 6] and Umberto Eco [18] in a field named semiotics.

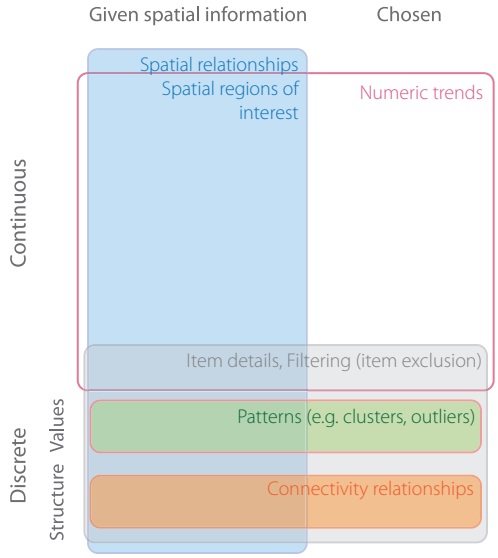

Fig. 10. Classification of visualization tasks. Tasks are defined based on whether the spatialization is given or chosen and if the design model is continuous or discrete. Adapted from [34]

Semiotics, or the study of signs and sign processes (semiosis) [36]. Semiotics as a field also encompasses how meaning is constructed and understood [36]. It is typically split up into 3 branches[36]:

- *semantics*: relation between signs and what they refer to (denotata);
- *syntactics*: relation among signs; and
- *pragmatics*: relation between signs and the effects they have on the people who use them.

Although semiotics is not confined solely to the visual system (it can also apply to meaning we give to scent)The application of semiotics in the context of visualization is particularly important when using visual metaphors to represent something. For instance, in cartography a cross has a meaning which happens to be a church (see *figure 10*. However, in a biological experiment workflow visualization, a cross may mean that an animal died. Within each field, the symbol means something different. Similary, a line on a map could indicate a road, however in the context of a biological experiment, a line can often be used to represent a DNA strand (see *figure 11*. These examples refer to the semantics of the sign.

In addition to this, we could use combinations of signs to create more complex structures with more meaning (e.g. this sentence), this refers to the syntactics. We could also organise these signs into taxonomies (where we have a hierarchy) to create a *visual taxonomy*, a small example of which is in *figure 12* where we arrange glyphs (visual entities representing information) in a hierarchical format based on classifications on data and material centric processes.

We can also use taxonomies to classify the properties of glyphs like those shown in figure 12. For instance, we may want to classify glyphs by the pre-attentive (things we take in about entities with paying much attention) and attentive (things we have to pay more attention to in order to take in its message) stimuli utilized in their creation. This has two major advantages:

- 1. it highlights the importance in considering such properties when creating glyphs; and
- 2. it would make it possible to search for glyphs based on their characteristics, e.g. shape, colour, etc.

Such a taxonomy has been created by Ropinski and Preim [28]. This taxonomy is summarised in figure 13.

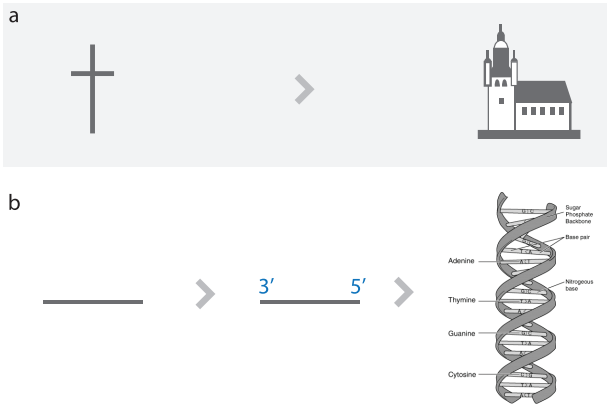

Fig. 11. The meaning of abstract images in differing fields, their *semantics*. a) a cross on a map has a meaning 'church'. b) a line in a biology themed visualization has the meaning 'DNA'. Use of these abstract structures to represent complex ones is key to cartography and glyph based visualizations in particular.

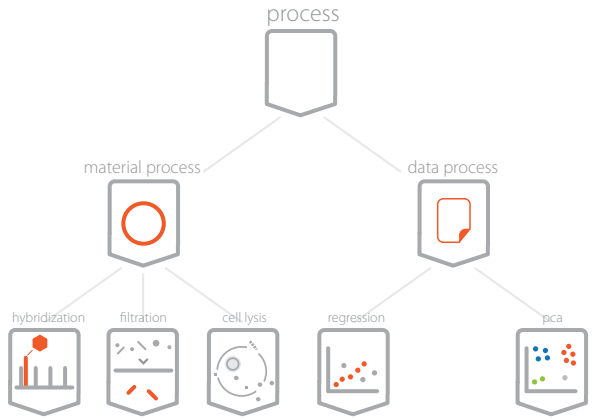

Fig. 12. Small example of a taxonomy for bio experiment processes utilising glyphs to represent information. We have an upper level process which has a certain shape and colour and have additional details inside as we go lower in the taxonomy. The process category has a constant outer shape.

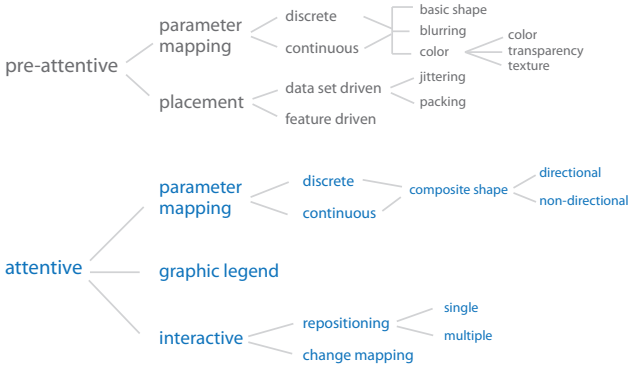

Fig. 13. Glyph taxonomy as presented in Ropinski and Preim [28].

3.3 Summary

The field of visualization taxonomy is much newer than say that of the biological domain but the benefits of having a taxonomy are to become clearly apparent when trying to navigate the growing visualization space. Some elements not covered by current taxonomies though is the definition of data, information and knowledge as discussed by Chen *et al* [12]. This is particularly important since the majority of taxonomies discussed in this section are providing classifications on data, the definition of which is interchangeable between fields. Future taxonomies should attempt to provide a definition of these terms.

There has been good progress in the development of a visualization taxonomy, and this will improve more given the development of visualization ontologies like that presented in Shu *et al* 2006 [30] and the continued rise of the semantic web and the linked data vision. It will therefore be important to ensure that descriptions of visualization tools are exposed in a semantically meaningful way. Moreover, being able to describe how visualizations were created (in the processes invoked) would allow for reproducibility of visualization tasks or could inform users better about the data rendered for them. For example, it is typical for visualizations to be presented to users but masking out uncertainty in terms of statistical bias, error ranges and so forth. Through informing users in a standardised way about how a visualization came to be constructed, they can make more informed decisions about whether a particular result is good or not, or can search through visualizations created using particular statistical methods for instance.

4 Conclusion

Although taxonomies originated in the field of biology, their application is pervasive and they are used to categorize content on a broad scale. However, taxonomies are unlikely to ever be perfect for a number of reasons:

- 1. we are mapping continuous phenomena into discrete space. It is an unnatural mapping and will almost certainly not be perfect;
- 2. there are always exceptions to the norm; and
- 3. there is an inability to unanimously agree with consensus.

Yoon approached this subject of imperfection in the book *Naming Nature: The Clash Between Instinct and Science*[37] in which she also refers to the umwelt concept when it comes to constructing taxonomies. In this book she states that since we have an inbuilt system of classification for the world around *us* as humans, classifying everything, that within *our* environment scope and without will inevitably lead to unnatural and incomplete classifications. As an extreme example, when thinking about the olfactory (smell) system, humans are much less sensitive to smell than dogs[37]. Therefore the way dogs classify things in their environment will be inherently different to the classification of the same environment by humans. If we apply this to people from differing cultures or generations, we will not get the same extreme variance we would get with a dog, but there will be some differences in the classifications that we get of the same thing. This is inherently problematic when it comes to those who make the decisions about what makes a species and what doesn't or which classification is best for visualization taxonomies.

That being said however, modelling our world as close as possible to the real world and in a generic a way as possible will certainly be useful in guiding users and researchers in describing and navigating domains (biology, chemistry) in a consistent way, something of growing importance in anticipation of the semantic web.

References

[1] Chemical entities of biological interest. http://bioportal.bioontology.org/ontologies/46576/.
[2] Gene ontology. http://bioportal.bioontology.org/ontologies/1070/.
[3] Phylocode. http://www.ohiou.edu/phylocode.
[4] Olive. http://otal.umd.edu/Olive/, 1999.
[5] A. Abela. Visualization taxonomies. http://extremepresentation.typepad.com/blog/2008/06/visualization-taxonomies.html.
[6] J. Bertin. *Semiology of graphics:diagrams, networks, maps*. 1983.
[7] J. Bertin and M. Barbut. *Semiologies graphique*. Mouton, Paris, 1973.
[8] K. Brodlie. *Scientific visualization: techniques and applications*. Springer-Verlag, 1992.

[9] J. Caporaso, C. Lauber, E. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. Gordon, and R. Knight. Moving pictures of the human microbiome. *Genome Biology*, 12(5), 2011.

[10] S. Card, J. Mackinlay, and B. Shneiderman. *Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, 1999.

[11] S. Card and J. MacKinley. The structure of the information visualization design space. *Proc. IEEE Symposium on Information Visualization*, 1997.

[12] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. van Liere, L. K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and knowledge in visualization. *IEEE computer graphics and applications*, 29(1), 2009.

[13] E. H. Chi. A taxonomy of visualization techniques using the data state reference model. *Information Visualization. InfoVis*, pages 69–75, 2000.

[14] E. H. Chi and J. T. Riedl. An operator interaction framework for visualization systems. *Symposium on Information Visualization*, pages 63–70, 1998.

[15] G. O. Consortium. The gene ontology in 2010: extensions and refinements. *Nucleic acids research*, 38 (Database Issue):331–335, 2010.

[16] C. Daassi, L. Nigay, and M.-C. Fauvet. A taxonomy of temporal data visualization techniques. 2006.

[17] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 2008.

[18] U. Eco. *A theory of semiotics*. Indiana University Press, 1979.

[19] B. Edwards, C. Reddy, R. Camilli, C. Carmichael, K. Longnecker, and B. V. Mooy. Rapid microbial respiration of oil from the deepwater horizon spill in offshore surface waters of the gulf of mexico. *Environmental Research Letters*, 6(035301), 2011.

[20] D. Harper. Taxonomy.

[21] A. Helbig, A. Knox, D. Parkin, G. Sangster, and M. Collinson. Guidelines for assigning species rank. *Ibis*, 144:518–525, 2002.

[22] K. Kull. Jakob von uexkuèll: An introduction. *Semiotica*, 2001(134):1–59, 2001.

[23] M. Manktelow. History of taxonomy. *Lecture from Dept. of Systematic Biology, Uppsala University*, 2010.

[24] E. Mayr. What a species is and what is not? *Phil Sci*, 63:26–77, 1996.

[25] S. Meiri and G. Mace. New taxonomy and the origin of species. *PLoS Biol*, 5(7), 2007.

[26] L. Proctor. The human microbiome project in 2011 and beyond. *Cell host and microbe.*, 10(4):287–291, 2011.

[27] D. Roam. *The Back of the Napkin*. 2008.

[28] T. Ropinski and B. Preim. Taxonomy and usage guidelines for glyph-based medical visualization. *Proceedings of the 19th Conference on Simulation and Visualization (SimVis08)*, 2008.

[29] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualization. *Proceedings IEEE Workshop Visual Languages*, pages 336–343, 1996.

[30] G. Shu, N. Avis, and O. Rana. Investigating visualization ontologies. *Proceedings of the UK e-Science All Hands Conference 2006.*, pages 249–257, 2006.

[31] M. G. Simpson. *Plant Systematics*. Academic Press., 2nd edition, 2010.

[32] M. Taylor. What is cladistics? how reliable is it? http://www.miketaylor.org.uk/dino/faq/s-class/clad/index.html, 2003.

[33] L. Tilton. From aristotle to linnaeus: the history of taxonomy. http://davesgarden.com/guides/articles/view/2051/.

[34] M. Tory and T. Möller. Rethinking visualization: A high-level taxonomy. *IEEE Symposium on Information Visualization*, pages 151–158, October 2004.

[35] J. von Uexküll. An introduction to umwelt. *Semiotica*, 2001:107–110, 1998.

[36] Wikipedia. Semiotics. http://en.wikipedia.org/wiki/Semiotics.

[37] C. K. Yoon. *NAMING NATURE: The Clash Between Instinct and Science.* W. W. Norton and Company, 2010.