Zachary W Painter. Creation and maintenance of taxonomies and ontologies. A Master's Paper for the M.S. in L.S degree. April, 2013. 32 pages. Advisor: Rebecca Vargha

Homo sapiens are a species obsessed with the classification of objects. From the various jars of Mesopotamia, to the Forms of Plato, to the Dewey Decimal System, civilization has made great strides as a species to organize the world. The modern world allows people to organize large amounts of information like never before with unrivaled accuracy. However, people often struggle to create new taxonomies and ontologies to satisfy their need to sort objects, and often find it difficult to merge new or updated information into taxonomies designed for a previous era or set of information. This paper will examine how various people and organizations have dealt with these problems of ontologies and classifications, while providing recommendations for each scenario.

Headings:

Classification – Study & teaching

Information Retrieval

CREATION AND MAINTENANCE OF TAXONOMIES AND ONTOLOGIES

by
Zachary W Painter

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April 2013

Approved by

_____

Rebecca Vargha

Table of Contents

Introduction and Theory

Human beings are obsessed with classification. From entertainment collections, to the books found in libraries, to the items found in a commercial enterprise, humans have an extreme compulsion to organize and make sense of the world around us. In order to classify objects, there must be a consistent system of organizing these materials. This system of organization is defined as a taxonomy, and stores the essential information for the objects that are classified – in other words it is the domain of the knowledge stored within the system.

Once the domain for classifying objects is created then, the next step is to create a theory for how one thinks of the objects within that domain. That system is called an ontology. For example, a domain could be dogs, while an example of ontology is the different breeds of dogs. A cat or a table would not be a dog, and would not be classified in the domain to begin with. A collie would not be classified in the same ontology as a terrier.

So what occurs when new information is revealed, or a new invention that changes the classification of data? The virtual explosion of digital media in modern life has created new formats and structures for classification. Does one build a taxonomy from scratch, or try to fit this new information into an existing taxonomic structure? If the information is integrated into a new taxonomic structure, then does one create a new set of ontologies to classify the data, or can an attempt be made to fit the information into a preexisting ontological set? This research paper contains historical analysis and interviews of persons and institutions that have dealt with the problems of classification

of new information, be it a new taxonomy or a new ontology, and the decisions that were made through the process.

The first step in creation of a taxonomy or ontology is to determine how to classify objects from a contextual view. Philosophy provides two paths as to how one perceives reality. For the ancient Greek philosopher Plato, universals are real. For the Scottish philosopher David Hume, particulars are real.

In many of his works – most notably *Phaedo* and the *Republic* – Plato references a concept known as "The Forms", which provide an understanding of the world and the things in it. All knowledge exists independently, outside of the physical world, in an intuitive space. This knowledge is classified into "The Forms", or the essence of the objects that we perceive. These are "Universals"; there is a universal, perfect, ideal template for an object, such as a chair or triangle, which can be perceived. The challenge is that one does not truly know where this abstract concept lies in reality.

In *An Inquiry Concerning Human Understanding*, David Hume challenges the notion of universals. While Plato would assert that people intuitively know what a chair looks like, Hume would deconstruct that to say that one truly does not know what the essence of a chair is without first using it. These are 'Particulars"; people only see specific instances of objects, rather than a universal template of objects, when looking at something like a chair. The problem with particulars is that if one cannot surely classify something as one thing, developing an accurate domain could be impossible.

Those who wish to create a taxonomy or ontology must grapple with the consequences of this debate. Those who favor universals will often have a system better suited for a broad overview of many different fields and subjects, while those who favor

particulars will often be superior at designing a system to differentiate minute details between similar objects.

Just as the works of Plato and Hume can be translated into a framework, the works of thinkers in other fields can be used in helping to understand taxonomies. Carl von Clausewitz, a Prussian soldier and military scientist, wrote in his seminal work, *On War*, about the concept of the "coup d'oeuil", or the blink of an eye:

> When all is said and done, it really is the commander's coup d'œil, his ability to see things simply, to identify the whole business of war completely with himself, that is the essence of good generalship. Only if the mind works in this comprehensive fashion can it achieve the freedom it needs to dominate events and not be dominated by them.

With one glance, or blink of an eye, a superior military leader should be able to assess the situation at hand and plan courses of action or responses without hesitation. Such speed of thought and action grants a higher chance of success for the military leader, as he can dictate the course of action, seize the initiative, concentrate and control his efforts, and act with surprise and speed. When pitted against a force of equal or superior ability, this ability to glance at the situation and make the correct decisions quickly is the only reliable way to ensure a favorable outcome.

On a similar note, a taxonomic classification should allow for the user to cast a glance of the eye and plan courses of action without hesitation. What sort of information is located in this taxonomy, and where is it located? Does one need a detailed familiarity with the system, or can a complete novice understand how the system works? Of course, a complete novice will need some assistance in most taxonomies, due to the nature of subject specialization and knowledge, but in general a simpler taxonomy is preferable for both the novice and the experienced user alike. A good taxonomy will be able to grant its

user enough descriptive capability to find what they want, while being simple enough to process and remember.

When asked to think about an example of a large taxonomy, most people will choose something from the sciences. Well established taxonomies are prevalent in all of the sciences for the classification of knowledge within that science. While these taxonomies and ontologies are often well understood and documented as exemplified with the UNC Herbarium, the initial understanding and creation of these systems was fraught with great difficulty.

> Correct understanding of early geological printed works and manuscripts can be difficult to achieve when lithological and mineralogical terminology unfamiliar to the reader is used. This is a prevalent problem in early texts when mineralogical and lithological names had not become standardized; in many cases names were descriptive in terms of the mineral habit and constituents (e.g. ruin agate, roe stone, or copper foam).

This poses two problems. First, a user without a background in a field will have difficulty understanding the appropriate terminology to classify objects, and objects that are not given a standardized name – particularly a problem for systems in the early phases of development – will be difficult to identify and categorize. As illustrated later in this paper at IntraHealth International, lack of standardization within ontologies can be very problematic for easily organizing and retrieving information.

Standardization is often a problem for institutions that are not fully professionalized, and the English have historically had a strong tradition of amateurism in their scientific pursuits. The great strengths of amateurism are that the low barrier of entry into the field allows for innovation and creativity to come from almost any source without the need for formal training or great financial resources. This freedom often comes with a high price; amateurism often means that each individual performs the

process in a different manner and that can lead to dangerous situations and inability to replicate or understand potentially significant findings.

In contrast, professionalization means formalized training and standardization of methods. This can limit the creative edge of individuals in a field and restrict access to only those who can afford training and materials. The result is that the standardization of practice will allow for more efficient research and conventions. In the world of taxonomies and ontologies, this means that a professional environment with as many participating members as possible will often create the best system, even if it means that such a system takes a longer time to develop due to the nature of compromise and research.

The key for those in data management is to standardize classification of naming conventions as soon as possible. Failure to do so can lead to problems such as translating information, vastly differing classification systems, or problems with integration of old or legacy information. Often, this means professionalization of the field that the system will be used in, or a professionalization of those assigned to create a classification system. Most scientific taxonomies and ontologies are now well established with little change, and this has been a process that has taken decades to perfectly create under the direction of increasing professionalism in the field. Emerging fields do not have this history and may not have the same trained professionalization to come up with a set of ideal standards, and will need to aim for standardization as soon as possible.

Of course, there will be situations when an organization will have to develop their own taxonomy internally to suit their specific needs. In these situations, it is likely that the new system will be developed using principles from a preexisting system, and that the

preexisting system will have been constructed by professionals. The scope of this research includes specific examples of institutions that have done so, and with success.

Interviews and Examples

Georg Wilhelm Steller traveled with Vitus Bering on his second voyage from Kamchatka to the American Pacific Northwest in 1741-1742. Steller was a German who had come to Russia seeking employment in 1734, and he would soon find it with the Russian Monarchy. Steller - a naturalist - kept a journal of the voyage and became the first trained naturalist to travel to the North Pacific. He made a number of unique observations that no other naturalist before or since has made, and his contributions to science are great. Steller's observations were mostly about the environment and the natural world he saw, but there were a number of comments on the crew and their actions. Of particular note are Steller’s comments regarding the objects he encountered on his travels.

Steller's first contact with native settlement occurred on 20 July, 1741. Steller and his Cossack assistant went ashore on Kayak Island to collect fresh water for the crew and to obtain samples of the environment for study. While in their journey, a native settlement was found. Steller took a few of the goods with him back to the ship for inspection, and the area was plundered. A few European goods were left in exchange, but this is of little consequence as the Natives would not have taken kindly to such an exchange. Every tool or object that Steller brought with him to observe was traced back to a European heritage; he could only envision the objects that he had previously seen

before and the purposes they had. Steller had difficulty of seeing the true process of the tool-making or the purpose of the tools without seeing their use by a native.

In the absence of such experience, Steller would have to cast his mind back to similar European tools and their uses. While they may have looked the same, European tools would have undoubtedly have had a different purpose than the American ones, and those ideas would not align to the actual ideas that the Americans would have had about the same tools. Not understanding the purpose of the tools would lead to an ignorance of its true use.

Steller also had an issue with the identification of many of the flora and fauna of the geographical area. Steller would attempt to identify any new thing he saw in the guise of the impressions he already had about these animals. Steller reports of seeing a “Sea Monkey” on 10 August 1741, a strange-looking water mammal that had a fairly unique behavior and an appearance that was unknown to Steller. Steller did his best to describe the creature and its habits, but was unable to come to a definitive conclusion about the true nature of it. This problem is not to be underestimated.

Steller was thinking within the cultural imprinting of his era, and his discovery would need to be linked to something that previously existed in European science. It would be impossible for Steller to simply conjure something that did not have his own cultural impression on it, and Steller would have been at the mercy of his experiential knowledge as to identification of the creature of the tools. There is no native culture that Steller would have experienced in a sufficient depth at this point to see these objects through another set of eyes – and thus no ability to contrast – but that lack of viewpoint is what keeps Steller locked on his own culture and ideas. It could perhaps be explained that

too much of a European viewpoint and focus locked Steller into this narrow sense of culture and identification. He could not think outside of the cultural framework that he has grown up with, and therefore cannot see the world and its objects like the natives of the Americas.

Comparison is really the only way of understanding such phenomenona. Steller calls the strange creature a monkey because he does not know of any other way to describe it. Steller lacks the ability to give the creature a new name because there is no other experience like it. Does experience truly preclude understanding? Experience certainly does preclude understanding; simply put, you cannot do what you do not know. In this case, Steller is attempting to name a creature that he has never seen before. Can he do this? The answer is no, because he has no such experience with the creature previously. Instead, Steller relates his experience to something that he previously knows. Comparison of the new idea to something that has previously been experienced is the only way for Steller to understand the new concept. By calling the creature a monkey, Steller remains wedded to his experience. It is difficult to simply imagine something brand new; one must first have an idea in place that can relate to the new idea itself. Such a problem with new schema could reduce a culture to a static entity at the most extreme point, but this is certainly not the case.

What that means for people involved in the world of data management and libraries is that collective past experience will shape the ways that objects are perceived. Staff will be unable to truly understand an object that they have never perceived before, and only their own perceptions of similar objects will assist in creating a classification. If using a preexisting taxonomy, they are likely to shoehorn the new object into a

preexisting ontology that may not be the best descriptor of the new object. In creating a new taxonomy or ontology, staff are likely to use concepts from systems that they are familiar with in order to build the new system. This could prove problematic if the old system does not truly align with the new need for description of an object.

Being descriptive and simple is a daunting task, and for the athletes and coaches of the game of American football; designing a taxonomic system for calling plays is often a large challenge. Some systems of play allow for decisions to be made after the ball is snapped, but these systems require immense time and energy to perfect and are rarely used above the small-college level. Most offensive and defensive systems require a very descriptive terminology to tell each player what they are doing, and these plays are often called or swapped over 75 times a game on both sides of the ball at high levels.

The researcher spoke with three current and former coaches, with experience ranging from the middle school level to the professional level, about their experiences with offensive play calling and numerology, or their reactions to offensive play calling.

Here is an example of the great differences between taxonomic systems: Jon Gruden, a former American professional coach and current analyst for ESPN and quarterback mentor, had a famous exchange with Auburn Tigers (college) quarterback Cam Newton before he was drafted as the first overall pick by the Carolina Panthers (professional) regarding verbiage and play calling. In one of his sessions, Gruden recited a play call for Newton to decipher that would likely be used in a pro-environment - “Flip Right Double-X Jet Counter 36 Naked Waggle 7-X Quarter”- and Newton was at a complete loss.

When asked to name a play that Auburn would call in the huddle, Newton replied that Auburn did not huddle, and that just seeing “36” on a board at the sideline was enough to let him call the play. This was stunning to Gruden, and he raised concerns about Newton’s ability to understand the workings of a complex professional system. Newton’s team had won the college football championship that year, which leads one to believe that unless Auburn was filled with transcendental players, the system that they used to call plays worked well enough and was descriptive enough.

One of the interviewed coaches is the creator of an offensive scheme known as the “Triple Shoot”, an offense which blends principles of the “Run and Shoot” and the “Flexbone Option” offense, as well as a few other influences. The Run and Shoot and Flexbone are both among the rare systems that decide how a play will be executed after the snap, but they accomplish their goals through wildly different means. The Run and Shoot is an offense that throws the ball with complex timing and route patterns for the receivers, while the Option is an offense that runs the ball with complex blocking schemes and decision making for the rushers. The nomenclature for both is wildly different, but by combining the two systems an offense could conceivably take an advantage over any defense due to diversification and uniqueness.

In order to combine his plays into one system, the coach needed to determine which plays fit together in a “series” or ontology. While this was made more difficult by the fact that the coach used two wildly different systems, thus creating problems with integrating plays from both systems together, it also gave him the chance to determine which plays from each offense he could keep and which ones he would have to discard. Because all offenses have core plays – or the plays they run most frequently – and

constraint plays – plays that counter the defensive tendency to stop core plays – the coach decided to use more of the core plays from each system while changing the order of plays so that he could have constraint plays to complement the true core of the system.

Another coach is the creator of a prominent coaching blog, and a high school offensive coach. He and I spoke about the naming conventions of plays and how the modern trend of playing without a huddle has simplified play calling. In particular, he referenced an article he wrote about the New England Patriots professional team. New England use an offensive terminology designed to be both descriptive and simple, and such a system allows them to execute at a very high level with an alarming speed of movement. New England has been among the league leaders in points, yards, plays run, and many other offensive categories because of the success of their offensive terminology. In fact, he thinks other professional teams cannot move at the same pace because they call plays like Jon Gruden, which is a slow method.

The third coach, a defensive backs coach and defensive coordinator, has remarked on this principle of speed and simplicity. Defensive play calling is fundamentally a different art from offensive play calling, yet the speed at which an offense moves will dramatically alter how he is able to construct a play on defense. Defenses often have a base setup, or a standard formation that they play against most opponents. While they are good general systems, they often are at a disadvantage against more specific offenses. Base defenses prefer to substitute to react to these offenses, but they cannot substitute against a fast paced offense that does not huddle. Teams like New England, Auburn, and countless other college and high schools do this so that they can take advantage of poorly constructed defenses. This speed is only possible with a classification system designed to

be executed quickly while being descriptive enough to allow the players the ability to process information quickly – a key element in the design of a taxonomy or ontology.

Cairo, Egypt, is home to two institutions that have problems classifying and organizing information in taxonomic structures. The Institut dominicain d'études orientales (Dominican Institute for Oriental Studies – Cairo) is a Catholic research institution devoted to the study of Islamic and Arabic culture. It was founded by the Dominican Order for purposes of scholarship and education, and serves as one of the chief centers of Islamic and Arabic study in Egypt. Also in Cairo is the "American University in Cairo" (AUC), an American-style, English-language liberal arts college.

In speaking with one professor and two doctoral students from the School of Information and Library Science at the University of North Carolina, this researcher had conversations about the challenges that these and other institutions in the Middle East face when building out their libraries and research centers. The professor has a background in the study of the Middle East and information trends in that region. Both doctoral students worked at the AUC and also have experience with IDEO-Cairo and other Middle Eastern libraries.

AUC is a primarily English speaking institution; both the AUC and IDEO-Cairo conduct operations and have patrons who speak in three languages – English, French, and Arabic. The United States Defense Language Institute is also another example of working in an environment with multiple languages which creates distinct challenges with classification and organization of material. Both institutions have adopted different approaches in confronting this problem.

IDEO-Cairo has attempted to solve the problem by hiring an external company from the United Kingdom to handle all of the classification of scholarly materials in their collection. They develop the taxonomy and ontology for the materials based on systems that they have designed, and then they give that system to IDEO-Cairo to use. The perception of librarians in the Middle East is often seen as clerical workers rather than full professionals, and having a third-party take over aspects of a library that require formalized and professional training is a way to sidestep the lack of trained workers. In addition, this allows the staff at IDEO-Cairo to keep the highest possible count of scholars and academics for research, rather than using personnel funds on highly-trained library staff. Since IDEO-Cairo is a small organization, saving costs whenever possible to maintain a high standard of quality in their work – in this case, Arabic and Islamic studies – is important for the financial and overall health of the organization.

The impact of this approach is that IDEO-Cairo does not have much control in how their collections are organized aside from telling the contractor what they would like to see, and there is a danger than no one on the IDEO-Cairo staff will know how to address a problem with classification when the need arises. For example, something might not be classified in a second section of interest, making it harder for the work to be discovered.

The AUC has taken a different approach by restricting the core languages down to one primary language for most of their resources and facilities– English. This action has reduced the need to have equal amounts of works in multiple languages, although the institute does collect in other languages – particularly in their Arabic Language Instruction Institute. Therefore multiple copies of works in various languages are not

necessary and allow the AUC to streamline their search processes and cataloging needs. Since one language dominates over the others, standards can be adapted to fit the needs of the dominant language without sacrificing quality in that language.

This focus on English comes at a higher price for the AUC. Because they are focused on English in an environment where English is second to another language (Arabic) and level with yet another (French), there is a risk that users who do not have a strong academic background in English may have difficulty categorizing and retrieving information. This limits the ability of the AUC to hire professionals who do not have an excellent command of English, and those workers are further handicapped due to the need to know Arabic and/or French in their daily lives around Cairo. The language limit for their staff can impact the professionalization of workers and potentially overload a few staff that makes the decisions on how to classify materials.

Both AUC and IDEO-Cairo have an advantage over many other modern academic institutions in the Middle East in that they have clear rules and a long history of scholarship. There are other institutions in the Middle East that have started rapidly expanding their collections in the past few years, and they often are facing a number of challenges in terms of classifying their information. These new institutions do not have to deal with the problems outmoded or discarded legacy data, and are truly building a resource collection from scratch.

The problem that these institutions have is the same as both IDEO-Cairo and AUC – they have difficulties when materials are not in their preferred target languages, and they often have to deal with the challenges of creating a brand new taxonomic system for classification without a base to build on. If the chosen system is missing a crucial

element, it may be difficult to go back and change the system. Both of those patterns emerge at the US Defense Language Institute Foreign Language Center and the US Environmental Protection Agency.

The United States Defense Language Institute Foreign Language Center, in Monterrey, California, is responsible for the non-English language instruction of military and Department of Defense government personnel. When a military officer needs to learn German for deployment in Central Europe, or if the Pentagon decides that security analysts should learn Arabic or Chinese as opposed to Russian, the DLIFLC is chiefly responsible for the instruction of those languages to US Department of Defense and government personnel.

This researcher spoke with a top librarian for the DLIFLC tasked with the oversight of library operations at the DLIFLC, and who makes decisions on policies of the library. The Aiso Library at the DLIFLC houses the print and digital media collections. With approximately 115,000 items in 39 separate foreign language collections, the Aiso Library is the information hub of a multi-ethnic, multi-cultural, and multi-lingual environment.

Her biggest challenge is that she only speaks one language to the level that is required for high level academic work – English. If she has a work in French or Spanish she can usually do a satisfactory job with translating and cataloging the material properly, but a language like Japanese or Kazakh is a not something she can do herself. She has a small library staff, but combined they do not speak every language that is offered at the DLIFLC and most of them do not perform both cataloging – the art of organizing where the material is located – and reference – the art of finding where the material is located.

For some languages and requests, locating or organizing materials to fit the diverse environment is impossible without devising an in-house system to meet those challenges.

The DLIFLC is also tasked with creating a specialized cataloging system for classifying their materials. The Anglo-American Cataloging Rules were the premier system of assigning materials to a specific ontological set.

However, the Anglo-American Cataloging Rules were designed to classify materials in the language of the home countries that designed it – English. Materials in other Indo-European languages, particularly in the Germanic or Romance families, could possibly be cataloged with few difficulties. However, materials in non-Indo-European languages, such as Arabic or Japanese, will often have formatting that does not align with the rules of AACR.

What she has done to work around the problem is to utilize the teaching staff to translate works for her. Some works that are in a language such as Uzbek might only be available in Uzbek, Kazakh, and Russian, so finding an instructor in either of those languages will assist in overcoming translation barriers into English. While these staff do not have the same specialization as she does with bibliographic recording and cataloging, she instructs them to find the important items in whatever work she needs cataloged and translate for her.

To avoid the problems inherent in the formatting of classification in other languages versus English and AACR, a fairly generic and universal method was constructed for organizing where information is placed on a shelf and in a digital archive. The needed information for accurate classification was reduced to the bare minimum needed to have unique and easily identifiable records – with the document type and some

sort of primary key (ISBN, ISSN, etc) forming the core of the record. This strategy allows a bypass of issues which are inherent in translating languages and get a clear record of what resources.

If a unique identifier is not available, the translator is asked for as much information as possible in the hopes that something is unique enough to distinguish that work, but relatable enough so that linking works with multiple languages can be noted for ease of searching. It is not a foolproof method, as some unique identifiers are difficult to come by, but by sorting materials by language first most of the problems arising from duplicate records can be avoided generally.

The collections at the DLIFLC are large and can rapidly change if world events were to change. For example, the decline of the Soviet Union and the rise of China and the Middle East have seen a shift away from languages like Russian and other Slavic and Caucasian languages, and a move towards Asiatic and Semitic languages. Therefore, the staff at the DFIFLC has decided to arrange their materials by common language, and they have sorted the languages in alphabetical order rather than language groups. The Germanic languages, such as Danish, English, German, Norwegian, and Yiddish would appear in that order on the shelves rather than family and subfamily grouping (for example, Danish and Norwegian would be placed near each other, with German and Yiddish near each other, and so forth), with the other languages such as French or Farsi placed in-between them according to alphabetical order in English.

While this allows for classification to easily start with the call tag of the language, thus making searches easier because all materials of that language are in the same space, it does an inadequate job of showing all related resources in other languages. In order to

do so, the system would have to be rearranged, which may take too much time for no return on investment. Systems that do allow this linking have an advantage for retrieval over those that do not, and that is present in the next example.

The United States Environmental Protection Agency is a government agency responsible for the protection of human health and the environment. There are EPA facilities throughout the United States to serve the mission of the agency. The EPA Office of Research and Development is the scientific research branch of the organization, tasked with providing a scientific basis for all of the other activities and policies by the organization.

The researcher interviewed a Program Analyst at for the Information Management Support Division of the Office of Science Information Management within the EPA ORD. He and his team of student contractors have developed a database for the storage of scientific publications within the Office of Research and Development at EPA. In conversation with the team, this researcher learned about the creation of this database – named VIVO – which uses a semantic data model known as RDF (Resource Description Framework), and the challenges that were evident in the creation of the taxonomy and ontology for the database.

Publications were pulled from the Thompson Reuters Web of Science database and uploaded into VIVO. Information about each scientist, from the journals that they have published in, their pay grade, their educational background, and other important pieces of information are all included in this database. While the IMSD staff are the only people working on this database now, the true goal is for any scientist at the EPA's Office of Research and Development to be able to add and edit data. Due to the nature of

RDF, all of the data in the system can be semantically linked up for ease of searching, and individuals can create API applications to exploit VIVO to better fit their needs. This process is more intuitive for the user and allows for easy linking of related pieces of information for faster, more efficient, and more in-depth queries.

The ability of VIVO to utilize an API platform is perhaps one of the greatest tools for a data manger in creating a system of organization. VIVO is a versatile tool that will let anyone say anything about anything. This is very useful for the user as they can correct data and make assertions about the knowledge that is in the system, but this comes with the price that a user can add in information that is incorrect. If the ORD were to decide that no one other than the IMSD office could add or change data in the system, then the user would lose the ability to manipulate the data to suit their needs or improve accuracy. Loss of such abilities or freedom could deter users from working with VIVO, which could lead to either the downgrade or cancellation of VIVO services. Such a result would be ineffective for the EPA.

Because no one at the EPA can say for certain about any publication activity that occurs outside of the EPA, the only people in the VIVO database are from the EPA. If someone wants to know about authors who are not in the EPA, they could not use VIVO without the assistance of an API unless they wanted to utilize the Web of Science, which is not intuitive and does not have the same semantic properties as VIVO. Web of Science also could be confusing or difficult for a user who does not have training or familiarity with reference or database software, and might be unhelpful for the scientist who decides to use that platform instead of VIVO. The utilization of an API, either built by EPA scientists, EPA information professionals, or a third-party of some sort, will be the

answer for those who want to freely edit VIVO with information that is relevant to their interests while maintaining a system that is useable by all.

The EPA has a number of titles and names for employees and organizations that do not match the names found in most database ontology sets. Because there is no preexisting ontology set to classify these titles, the information staff at the EPA had to create an ontology set to place into VIVO to organize this information. Such an ontology would have to mesh with the other ontologies in the system while being different enough to warrant a brand new ontology. In this instance, it would not be possible to shoehorn the names and titles needed into an existing ontology.

VIVO uses a controlled vocabulary in order to standardize the information in the system, which makes it easier to locate and retrieve. The controlled vocabulary can come from a number of ontology sets, and the EPA uses several. The VIVO core ontology forms the base of the list, and from there common sets such as Dublin Core, Friend-of-a-Friend (FOAF), and Functional Requirements for Bibliographic Records (FRBR) are used, as well as others specialized for work in the sciences and government. From there, the parent institution can create ontologies for their specific needs.

The EPA:VIVO ontology was developed with these limitations in mind, and has fleshed out the information in the VIVO database by providing the needed specificity to enter in things like wage scales, office locations, and employee titles. To create this ontology, the IMSD staff detailed everything they would want to know about an employee at EPA, and then used those details to create an ontology class for each item, using other ontological classifications as a guide in construction.

While the EPA has a large amount of legacy data, they are filtering it out of their system before it becomes live for the users. In contrast, a company known as IntraHealth International has not done so with their latest organizational system, and it has caused issues in information retrieval.

IntraHealth International is a nonprofit organization based out of Chapel Hill, North Carolina, which promotes health and health workers across the world – especially in the underdeveloped third-world. IHI partners with government agencies, other nonprofits, aid organizations, medical institutions, and charity/patron groups to accomplish its goals of providing adequate health and care to people in the developing world who lack the resources or means to do so internally.

The researcher interviewed a Knowledge Manager and Resource Officer for the IntraHealth Chapel Hill Office. Her responsibilities include the SharePoint systems as well as the on-site library, which features a digital collection and a physical collection. In terms of this case study and the challenges of running two of the systems – the library cataloging system and the description system of the IHI digital archive (known as the Hive). The former is tailored to a specific environment, while the latter has flaws with legacy data.

IHI uses a special library cataloging system to sort their materials. While there are many different standardized cataloging systems – the most utilized being the Library of Congress system and the Dewey Decimal system – The IHI Staff created their own classification system. IHI has a specialized population of materials and staff. Using a traditional library classification system would be inefficient, as a very wide range of call

numbers could be used, and most of the materials might be classified under the same broad classification.

The creation of the new taxonomy for classification was based on looking through all of the resources at IHI and figuring out the appropriate ontologies to sort them under. Fortunately for IHI, this was an easy process once they determined what ontology set to use. IHI has Technical Areas, or specialized areas for their projects. Health Workforce Management, Malaria, Training, and Family Planning are examples of IHI Technical Areas. IHI is active in many countries across the world and has extensive cultural resources for interacting in all of those environments, so there was a need to place those resources within another call system.

The collection is big enough that a detailed system is required, but small enough to allow the call numbers to be lumped together to create a more user-friendly system. The decision was made to give each technical area a number, and then assign other related resources a number depending on what field that they were in. The base numbers were arranged from 1 to 9, with one or two decimal places to delineate subfields. If the materials corresponded to an IHI project, then an abbreviation of that project would be used in the beginning of the call number. Two letters marked the name of the author, and the year was added to the back. This created a number that was very descriptive, which is good for the user, while being very simple and easily fit onto either a screen or dusk jacket, which is also good for the user.

The Hive is the digital archive of IHI products and internal reports. It serves as an online catalog and information management system for all of the products that IHI produces. In order for the Hive to be successful, materials need to be cataloged as

accurately as possible. One of the critical tags for searching materials is the subject area. When the resource team at IHI migrated their legacy data into the Hive, they brought with them nearly forty years of material. Prior to the Hive, users who submitted a resource for the center could use any subject tags that they wanted. This led to several challenges with the classification of data. For example:

- Two fields were labeled under Abstinence. The first was “Abstinence” and the second was “Abstinence, Be Faithful, Condom Use”. Only one resource was filed under the second label.
- Contraceptives had a number of fields, but so did other forms of birth control. Condoms, IUDs, the Pill, and other contraceptives were given their own subject headings, which balkanized the search results for contraceptives.
- Husband-Wife Communication, Partner Communication, Marriage Communication, and Relationship Communication were all given individual fields.

For a user wishing to accurately find information, retrieving information with legacy data as outlined above were extremely problematic. The KM staff at IHI have recently begun to address this problem, foremost by restricting rights to adding materials to the collection to KM staff only. This is in contrast to the aim for VIVO at the EPA who want the users to be able to edit data, but the chief difference is that the EPA have placed heavy restrictions on their products before making the system live, and do not have to clean up messy legacy data. For a company like IHI, who have to update their data, restriction of edit privileges to appropriate staff only is a necessary step in creating a clean system.

Secondly, the KM staff have undertaken a project to modernize and update the vocabulary by removing misspellings (like “agrigultural development”, which had one resource under it compared to twenty-nine for “agricultural development”) and combining fields that have one or two resources with fields that are larger. This will

allow the user to more easily navigate the search box with more accurate subject headings while providing enough breadth and depth to narrow searches into a very specific focus for best results.

IntraHealth has shown that it is possible to develop a specialized system of classification for internal purposes, but legacy data is always an issue. While IHI has a large amount of legacy data, their records pale in comparison to the UNC-Chapel Hill Herbarium, who face the challenge of using an outdated system for their information.

The University of North Carolina at Chapel Hill maintains one of the largest and most prominent herbaria in the Southeastern United States, with over 750,000 plant specimens dating to 1835. Organizationally, it is a part of the North Carolina Botanical Garden, and houses most of the historical and research tools for the plants in the Garden.

The researcher interviewed a botanist and curator of the UNC Herbarium. In conversation with him, this researcher talked about the challenges of updating a large legacy collection and the effects of modern relational database technology and big data in the field of biology. In particular, he explained how the modern era has made cataloging and sorting information easier than ever before, especially for individuals who are trained in the subject of the taxonomy. There was also a discussion regarding changes in information and the impact of the development and creation of systems of classification, both at the taxonomy level and the ontology level.

With over 750,000 species of plants from across North Carolina, the Southeast, and the World, any movement of various plant families, genera, or species could require an extensive modification to the way that the plants are organized. The biological taxonomy is pretty well established above this level (Kingdom, Class, etc), so the plants

can easily be sorted in levels above family. Most of the creation of taxonomic structures is done at the species level, particularly when very minute differences warrant the branching or stitching of species.

The UNC Herbarium uses a biological taxonomy that dates to the late 19th century. This taxonomy has been updated several times and is now very out of date. The Herbarium still uses this system and has no plans to update in the near future. There are a number of reasons for this, but chiefly the return on investment for moving these materials is not worth the time or effort. The collection is so large, and taxonomic classification changes just enough at minute levels that materials could have to move drastically due to space constraints.

Some materials are not properly classified within an old taxonomic system. Legacy data can complicate the historical records and leave many taxonomic records in an ambiguous state. There are many older records that do not have a strong historical line, and the only way to make sense of them is to have modern citations and data to fill the gaps. For example, most of the ferns are not even sorted at a specific level because the fern taxonomy is in a constant state of flux. What good is it to move an object when it will just be moved somewhere else in a few years? The Defense Language Institute suffers the same problem, and both the DLIFLC and UNC Herbarium have addressed it in roughly similar ways, even to the alphabetical classification of most of their materials as opposed to classification by similarities. Both have taken advantage of the modern relational database to filter and categorize their information, and both have been successful in using it.

Indeed, the wonders of big data and the modern relational database have given the UNC Herbarium a chance to decline moving their objects around. As an example, the botanist has proposed that the genus *Marshallia* has a new, previously unrecorded species in *Marshallia legrandii*. Previously, *Marshallia legrandii* had been classified as *Marshallia obovata*, but he found enough unique traits in a few collected specimens to warrant the creation of a new species. This would not be a difficult move to make, given that both of these *Marshallia* species would be located in the same general area and would not need to move much, but splits like this are common enough to make record keeping difficult.

Another example from the collection is the traveling *Asteraceae*. This was one family of Angiosperms that were broken into three separate families, and then eventually stitched back into one family. The Herbarium staff initially moved the families to their new homes, which required a major renovation of cabinet space since the families would be places far apart. A few years later, when the family was pieced back together, the staff elected to keep the family split due to the hassle of moving everything again. The Herbarium does not have a large physical space to store their collections, and due to their policy of rarely weeding the collection, physical storage space is placed at a higher premium. With greater numbers of plant species collected every day, there may come a point where new brand species to the taxonomy are simply placed together rather than with their families due to the challenge of moving all of the materials.

It should be noted that few untrained members of the public come to investigate the Herbarium. Most of the users of the UNC collection are trained botanists who know what they are looking for and need little assistance. If it is not in the correct area of shelf

space, the database system will track it down exactly. Some items in the collection are loaned out, up to periods of ten years, but the Herbarium often has enough copies of the most requested materials to allow greater access and study. This means than many of the problems that would be expected of such a large collection using an outdated system of classification are nullified, since the technology can find anything that the highly trained users cannot. If the user base is well trained, and has a good working knowledge of the taxonomic system, minor updates and modifications diminish in importance.

Conclusions

Throughout this paper the case studies illustrate the challenges that organizations face when developing new taxonomies and ontologies. Both historically and in the modern era, in the fields of science, or the arts, and with large institutions or small, classifying and sorting information is filled with many issues. Below are the specific recommendations of this paper for creating new taxonomies and ontologies, or fitting new data into a preexisting system:

1. Whenever possible, use preexisting taxonomic and ontologic classification schema, preferably professionally developed, to ease the burden of creating a brand new schema.
2. When designing a new system, place appropriate restrictions on fields to edit and information that can be added, while giving users the ability to modify the system to suit their needs through personal applications.

3. Constantly take stock of legacy information and formats, and determine ways to streamline and update them when return on investment makes sense.

4. When designing a new system, keep the classification formats as clear and simple as possible, while retaining the ability to be descriptive enough for a user to find the information that they need.

5. Professionalization is critical to the success of developing a grand taxonomy, but it must be easy enough for an amateur to understand.

Bibliography

Chris Brown, "Speak My Language," *Grantland* (blog), January 16, 2013, http://www.grantland.com/story/_/id/8849439/how-terminology-erhardt-perkins-system-helped-maintain-dominance-tom-brady-patriots.

G.W. Steller, translated by F.A. Golder. *Bering's Voyages:Journal of the Sea Voyage from Kamchatka to America and Return on the Second Expedition*. 1741-1742.

Barbara Tversky. "Parts, partonomies, and taxonomies." *Developmental Psychology* 25, no. 6 (1989): 983-95.

Tversky. "Categories and parts." *Noun classes and categorization* (1985): 63-75.

Tversky, "Development of taxonomic organization of named and pictured categories." *Developmental Psychology* 21, no. 6 (1985): 1111-19.

Carl von Clausewitz, *On War*, (Princeton, NJ: Princeton University Press, 1984). Translated by Michael Howard and Peter Paret.

Alan Weakley, "A new species of Marshallia (Asteraceae, Helenieae, Marshalliinae) from mafic woodlands and barrens of North Carolina and Virginia.," *Phytoneuron*, 105 (2012): 1-17.

The Republic of Plato, Phaedo and David Hume's *An Inquiry Concerning Human Understanding* are taken from the chapters “Universals Are Real” and “Particulars Are Real”: Gould, James. *Classic Philosophical Questions*, (Upper Saddle River, NJ: Pearson, 2007), 335-350.

Personal Interviews conducted with the following: Three American football coaches, Four US Government Employees, Two UNC Professors, Two UNC Doctoral Students, and One NGO