Q1.

. del.icio.us and Memex

- Describe the purposes and mechanisms of del.icio.us using specific concepts and terms from Vannevar Bush's Memex. **(5 points)**

- Compare the del.icio.us mechanisms with the mechanisms for information management in Memex that have the same purpose. (You don't need to discuss any aspect of Memex that has no analogue in del.icio.us). **(5 points)**

A.

- del.icio.us is a social bookmarking tool that allows a user to bookmark—save for future reference or denote as particularly interesting—a link to a source on the Internet, just as the Memex acts as a repository for a user's information resources (plus the added bonus of notes, communications, and recorded conversations). The computer screen through which a user views del.icio.us can be thought of as similar to the Memex's translucent screens through which a user views all records. In order to add a bookmark to del.icio.us, a user can either manually enter the link in a field or add a button to their browser toolbar that, when clicked, will pop up a window and automatically fill in the information of the link location and title. But the Memex's most important asset is enabling association—that is, allowing a user to create a trail and "tie" resources together. del.icio.us enables this kind of trail-making with tagging. A user applies "tags" to the link with any number of key words or phrases that they feel accurately describe the article's subject matter or context, just as a Memex user "codes" material to be stored. A user can also create user-specific tags, such as "Myi202Blogs," in order to classify bookmarks as belonging to a similar category (an equivalence class), building a user-specific trail as in the highly individualized Memex. A Memex user can tape a single key to "join" resources, just as a del.icio.us can select an existing tag from their tag set or from a set of suggested tags in the collective del.icio.us user *tag*sonomy, thus adding them to a set of associated materials. Because Vannevar Bush had not foreseen the Internet, the Memex relies mainly on more analog indexing—microfilm that a user purchases and inserts into the desk containing the Memex mechanisms.

- del.icio.us is an online repository for articles and records found online that a user can access and retrieve information from rather quickly and easily, and is based on that user's highly individualized associations and categorization. The Memex, similarly, is a mechanized device that allows a user to access stored material with "speed and flexibility."

- del.icio.us, like the Memex, can be thought of as a "supplement to a user's memory" because it builds a highly personalized associative trail of resources for a user.
- Coding in the Memex, tagging in del.icio.us — codes a Memex user enters frequently are considered "mnemonic," similar to tags, which are completely biased based on a user's needs and indexing style. In both systems, codes are stored and easily recalled in order to view information categorized in the user's systematic (or unsystematic, in the cases of some online taggers) way.

The social nature of del.icio.us allows other users to see what items uploaded by others share tags, thus creating a social collective index of bookmarks and shared personal information trails. Bush foresees whole "encyclopedias" in the future that are composed of associative information trails created by other Memex users. del.icio.us enable digital knowledge sharing, while the Memex's promises are analog. Users can also pass their memex trails along to friends to share personal trail records, the way other users can view other users' trails in del.icio.us.

B.

1. terms used in memex: web of trails, associative indexing, improved microfilm, photography. Public and private viewing. Retrace steps, which allow for further investigation or intelligence.

The memex, first described by Vannevar Bush in his essay "As We May Think" in 1945, postulates a machine that looks like an ordinary desk, but which does extraordinary things. It holds an equivalent of a microphone, a camera, and an encyclopedia of information stored on microfiche. Bush imagines a researcher at this desk, reading and annotating (with camera, microphone, and by hand), creating a "web of trails" through various forms of information, allowing both the researcher or anyone else, to follow along at a later point. This web would mark both the path that the researcher took and the thought processes that accompanied him along it. Creating this "associative indexing" would allow researchers and scientists to quickly and relatively easily manage information in a given subject.

While Delicious was not designed to imitate the ideas described in Bush's essay, Delicious could be considered a modern-day form of this memex. It is a web-based service which allows bookmarking of websites and tags them according to a user-chosen set of words. It can created an associative indexing by linking together common tags and ideas allowing for easy retrieval and another way to identify commonalities of interests.

The two are also similar in terms of considering both public and private uses. In both cases, trails or tags are designed to be both useful to oneself, and to others following along.

However, Delicious does not create a “web” in terms of time. You cannot follow along a trail, whether of your own making or another. In this way it is dissimilar to Bush's memex. However, in the IO lab, we expanded on this idea, and students created different ways of creating a personal web trail that could be followed, showing that it could be done.

2. The mechanisms of information management in Bush's memex include storage of various kinds of media, which include voice, hand-written notes, and printed words embedded in microfiche. Ideas and concepts were linked together by the user, and could be retrieved at a later time, by either the user or an other party.

Delicious focuses more narrowly on web-based material, although the scope of the web encompasses many different media types, all of which can be easily bookmarked and tagged. However, the goal of the memex and Delicious is similar in terms of information management – easy retrieval and the discovery of associated concepts through the tagging mechanism.

-additional mechanisms in Delicious: public tagging, suggested tags, tag clouds

Q2.

Categorization and Bias

Explain each of the following ***two*** parts of a claim about the design of categories:

• "*Every system of categories is biased*…" **(5 points)**

• "*…and not every system of categories is equally good*" **(5 points)**

A.

Every system of categories is biased because the person (or organization) who devised that system is operating out of his/her own individual or cultural assumptions and way of looking at and naming things in the world. Further, categories are biased because they often serve the agenda of the one creating those categories.

An example of this kind of bias is the CAFE standard that classifies certain vehicles like the PT Cruiser as "light trucks" instead of "cars" so they don't have to follow the environmental regulations that "cars" are expected to follow. Environmental activists want vehicles to adhere to emissions standards, but of course automakers want to protect their own agenda and don't want to have to change their production protocols or standards.

Not every system of categories is equally good because that system may not actually describe its members well. That is to say, some system of categories may not actually reflect the external world as it is. Further, a system of categories may be "bad" in that it cannot be used to promote effective communication between different entities. Or it may be "bad" in that the user of that system (who has his own categories) cannot readily access the system's information because its categorization is so vastly different than his own.

For example, this ancient Chinese categorization of animals, as divided into:

(a) belonging to the Emperor, (b) embalmed, (c) tame, (d) suckling pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies.

would simply not be as good as a system like (a) mammals, (b) reptiles, etc. to help modern Americans with some task like searching for pictures of animals on an photo-sharing website. The categories we have for animals look nothing like the one above.

B.

All categories are constructed by an individual, organization, or society in response to that body's specific need. As a result, the categorization system the body develops will pay particular attention to the features most salient to the creator at the time of creation. An excellent example of this is the Dewey Decimal System (DDS). If you look at the DDS category for religion (200), it contains seven categories related to the bible and christianity, and only one category for “Other Religions.” There's a good reason for this – Dewey was categorizing the books he had available to him, which were primarily Judeo-Christian. As a result, his 'Religion' category reflects the state of Dewey's point of view about the world at the time he designed the system. This bias implicity expresses itself in a concept of what is “standard” (in this case Christianity) and what is not-standard (“Other Religions”).

It is easy to design a categorization system – simply pick any set of non-overlapping features about an object and classify the objects into categories based on those features. However, this is not necessarilly useful. I could theoretically organize books into categories based on their binding color, the weight of the paper they're written on, and the fruit the book most reminds me of. Yet this system would not be useful for the average user attempting to locate a book. Moreover, different users of the system might not agree on my categorization of a book (violation of objectivity). Good classifications contain categories that make sense to potential users of the system, are non-overlapping, are organized sensically (in the case of hierarchical categorization systems), and cover existing and future possible elements in the set of objects being classified. A system of categories that is well organized and logical will make retrieval (the ultimate goal of such a system) much easier than a system utilizing imprecise or idiosyncratic categories, or a system that is poorly structured and difficult to navigate.

Q3.

The Document Type Spectrum

• Why is it useful to think of document types as points on a Document Type Spectrum? **(6 points)**

• How does the mixture of presentation, content, and structure rules for document types differ across the Document Type Spectrum? **(4 points)**

A.

The Document Type spectrum offers an interesting view of various document types, since:

- Many aspects of document types seem to vary in a gradient-like nature.
    - This means that a strict classification theme is not the best view of document types.
    - Also, determining where documents lie on the spectrum can lead to information about how it must be handled.
- It allows us to extract common elements from documents, which enables design of better document management systems.

Some of the observations the spectrum gives us are:

- As document types move from being narrative to transactional in nature, the processing power needed to manage them reduces.
- The information in the documents themselves, is a lot easier to understand.

The content, presentation and structure rules for documents vary as follows:

- For very narrative documents (eg a novel), content and presentation are very closely integrated. Structure usually doesn't go beyond a table of contents, since these documents are not designed for people to "reference" information.
- As we move forward, documents like textbooks have the requirement as objects of learning, and hence have a clearer demarcation for content and presentation. Structure also becomes important – tables, indices etc and also more visible.
- Documents that are like dictionaries/encyclopaedias contain information in a very browser-oriented format. Content, presentation, and structure is clearly more separate.
- Web-based documents have an even clearer demarcation, and through use of structure, browsing them is a lot easier.

- The absolute transactional documents are well structured, and presentation is in a sense to expose facts.

B.

Often you encounter a distinction between “documents” and “data”, with different methods accompanying how we analyze both. However, most things feature a mix of the two, with a continuum of things that lie between pure documents and pure data. For example, a catalog page can contain both a prose description of how awesome a particular escalator model is in addition to more structured data about its price, model number and availability. We love to see tables and charts within an academic paper because that means less words we actually have to read. In the end, when we want to analyze either of these things, we do encounter many of the same problems and issues. For both thing we want to identify and separate the content, structure and presentation (although these components differ on the type of document and where it falls on the document type spectrum). We encounter the vocabulary problem in each of these, we need to identify and resolve issues with synonyms and homonyms. We want to identify and organize good content components to be able to assemble them into a useful, hierarchical document model. !

As mentioned above, the mixture of presentation, content and structure varies along the document type spectrum. At one end of the spectrum we have narrative document types (things like novels) and at the other end we have transactional types (things like databases). As narrative documents are authored by humans, they are highly designed with presentation in mind (unlike machines, we don’t deal well with pure data). A specific font and formatting may accompany the title of a paper, the abstract and the actual body, each signaling different structural components. As such, the data types and content within a narrative document are weakly typed, just plain text, and the overall structure is more heterogeneous. On the other hand transactional documents are authored by machines. The presentation is often quite arbitrary, which doesn’t matter as much because the data is strongly typed and structured.

Q4.

The “Abstraction Hierarchy of the Work”

An issue long discussed in information science is the “*abstraction hierarchy of the work*.”

- Briefly explain this abstract challenge. **(4 points)**

- Discuss why it remains an important part of the intellectual foundations of information organization, or else argue that it no longer matters very much. Use specific examples or “use cases” to make your points. **(6 points)**

A.

The abstraction hierarchy is a concept in which any information artifact can be described in a conceptual hierarchy that identifies its membership in a work. At the top of the abstraction spectrum is the concept of "work", which is an abstract intellectual or artistic creation, for instance "The Hitchhiker's Guide to the Galaxy" (which has taken form in radio shows, novels, games, and more). At the bottom of the spectrum is an item, or a single example of a manifestation of this work—like the specific copy of THHGTG held in a library. The challenge of abstraction is identifying the conceptual ancestry of any given artifact; identifying what attributes qualify the differentiation of a different classification of expression, etc.

The abstraction hierarchy is important for a variety of reasons. An important element of information organization is the intellectual classification of information, and doing so requires a clear understanding of the knowledge structure implicit in information artifacts. Practically, the abstraction hierarchy is a prerequisite for cataloging, because every item that a library (for example) catalogs must be identified as a particular manifestation of a particular expression of a particular work. Without the ability to create an abstraction hierarchy, classification would be impossible—that is, every item would be seen as unique and could not be retrieved according to the attributes of its intellectual parents. The abstraction hierarchy also plays a critical role in intellectual property concerns, as it's very important to identify the abstraction level at which intellectual property can be held and how related items can be used.

Q5.

Single-Sourcing

A company with a large collection of documents in word processing formats like Word or Word Perfect might prefer to have them encoded in XML so that it could employ a "*single source publishing strategy.*"

- Briefly explain "single sourcing." **(2 points)**
- What are the benefits of single sourcing? **(4 points)**
- What are the costs imposed by single sourcing? **(4 points)**

A.

A company with a large collection of documents in word processing formats like Word or Word Perfect might prefer to have them encoded in XML so that it could employ a "*single source publishing strategy.*"

- Briefly explain "single sourcing." **(2 points)**

"Single-source" is the attempt to create content once and use it in a variety of applications. As Ann Rockley demonstrated in "Managing Enterprise Content," a car company could write marketing material about a car once, and then that material could be repurposed for various needs such as brochures and websites, with differing amounts of detail based on the need. This eliminates redundant work and creates a single repository for an authoritative version.

- What are the benefits of single sourcing? **(4 points)**

  Single sourcing eliminates redundant work and creates a single repository for an authoritative version. This means that later users can all draw from the original author's work instead of creating new (and possibly contradictory versions). As the AGU discovered, single source publishing is particularly effective for their publishing process because they were able to create an authoritative XML version of an article and the transform it into PDF and HTML versions. The HTML version allowed them to publish their articles online and have the benefit of linked citations. By creating one original document that was able to be reused in various contexts, AGU successfully moved to single source publishing.

- What are the costs imposed by single sourcing? **(4 points)**

Like any information organization process, there is a cost involved. The tradeoff of doing more work in the creation of material pays of by being able to use the material in a variety of contexts, provided the original material was created in a way that took into account the later users' needs. In order to successfully achieve single-

sourcing, the original author needs to know all of the needs of the users downstream. This can be expensive to anticipate, and it can lead to unanticipated surprises by users who use the content later if the original materials are insufficient. It can require the original author to create much more extensive materials that he or she would typically create, and the model that the original author uses to create this material must be highly thought through so that it is complete and fully useable for all parties. This can require that a company change the way it does business, which can be a significant cost.

B.

Single Sourcing is the concept that a document should exist as a single master copy, which can simply be transformed and ported to other forms, viz. web, print, CD, etc.

More formally, this would mean:

- Using normalization to avoid problems caused by several copies of the same document
- Storing the "transformation process" which converts documents across different forms, as opposed to storing multiple copies.
- Using transforms to propagate changes/updates.

The key benefits are:

- Documents are stable since they are all controlled centrally (vs. individual control on each copy).
    - Fault isolation is a lot easier, since everything has the same origin.
- Consistency and Quality is better.
    - Acting on a single, updated master-copy ensures creation of the latest/error free editions of documents.
- Since the content is structured, it makes manipulations a lot easier.
    - Defining structural components of the document make it easier to manipulate.
- In practical terms, it improves productivity by having a controlled "system" as opposed to Ad-hoc processes.
- Moving to newer formats of documents becomes a lot easier.

The costs imposed by single sourcing are:

- Number of content components increase, and are smaller in size. In order to manipulate the information at a very granular level, it becomes essential to break it down into several parts. This increases processing costs.
- Policies and Practices:

    - The Single Source assumes a lot of importance, and hence a clear set of rules is required to determine the right procedures for access and control of the source.
    - The *enforcement* of these policies is another issue altogether.
- Organizational Changes:
    - A lot of changes need to be made to the way tasks are carried out.
    - People's roles change drastically too, and such a redesign is hard to implement.
- Backup/restore functions need to be more robust, due to the concentrated nature of the system.

Q6.

Authority Control

Authority control is a traditional topic in library science and remains an important issue in enterprise and inter-enterprise information management.

• What is authority control? **(2 points)**

• What problems does authority control attempt to solve? **(2 points)**

• Compare the methods and effectiveness of authority control in the library setting, inside single enterprise, and in inter-enterprise contexts. **(6 points)**

A.

Authority control deals with the creation and maintenance of a controlled vocabulary (its terms and what these refer to). The necessity for the maintenance of a controlled vocabulary arises from its artificial nature, needs and contexts may change over time. !

Authority control attempts to resolve ambiguity (the vocabulary problem) for things such as names, places and times. For example, under what authoritative name should a person be listed under? Myself, my full name is Franklin Yo-Shang Cheng, but my parents never call me Franklin at home, they call me Yo-Shang. I have Franklin Cheng or Franklin Y. Cheng on official documents like my driver's license and school transcripts. People call me Yo-Shang and my profile on the iSchool website has me under F. Yo-Shang Cheng. This is the name-matching problem or co-reference resolution. Can people and machines figure out that all of these names refer to me (sadly they often do not). The same issues arise with place names and times, where there may be variant spellings or transliterations (Beijing vs Peking), or if a term is just vague (Elizabethan Era). !

There are varying methods and effectiveness to authority control in the library, single enterprise and inter-enterprise environments. Libraries have dealt with this issue for a very long time and the people working on the problem are trained professionals. In particular the Library of Congress maintains an "authority file" for person names, titles, geographic areas, etc. With that said, even with trained professionals with a lot of knowledge about particular domains, things can escape their grasp, as we saw with Bob's multiple entries (Robert John Glushko vs Robert J. Glushko). For enterprises authority control arises when dealing with interoperability issues. A shared vocabulary and context is needed in order to exchange information, whether it be within an organization or between organizations. This gives rise to the creation of standards. Within an enterprise, it is difficult for a single body to know what the specific information needs of each department are. It is easy to set a particular mandate for the entire organization, but

these are never effective. There's a tradeoff between the complexity of a standard and the number of people who are able to agree on it. Between enterprises, smaller firms often have to bow down to the larger ones in order to do business together. If you want to deal with Walmart, you need to use their terms, since more is at stake for you. However, there is a lot of tension in having to agree on a shared language, as it means at least one party must invest in changing their information.

B.

Authority control is a traditional topic in library science and remains an important issue in enterprise and inter-enterprise information management.

• What is authority control? **(2 points)**

Authority control is the creation and maintenance of a controlled vocabulary based on a principled set of rules.

For example, the Library of Congress maintains an "authority file" in MARC format for the names of persons, corporate entities, geographic entities, and georgraphic names of political entities and titles of works. This authority file represents the standard for which these types of information are encoded.

• What problems does authority control attempt to solve? **(2 points)**

Authority control primarily attempts to solve the ubiquitous vocabulary problem, which, in the absence of a controlled vocabulary, many different variant of names may refer to the same thing or the same name may be referring to multiple different things. Authority control is also used to control variants of names resulting from misspellings, orthographical alternatives, language transliterations, nicknames, name permutations or obmissions.

For example, depending on the period of time or the particular usage, the capital of China may be referred to as Peking or Beijing.

• Compare the methods and effectiveness of authority control in the library setting, inside single enterprise, and in inter-enterprise contexts. **(6 points)**

In the library setting, authority control is implemented by standards such as MARC, which are widely used and generally less disputed. This is the most effective setting because the standards have been established and there are clear enforcement mechanims that can make practitioners adhere to the standards. In a single enterprise, authority controlled is enforced by enterprise standards and processes. Enforcement mechanims may be ineffective or the standards subjected to political disputes between silos. Therefore, the effectiveness is diminished but still useful. In the inter-enterprise contexts, authority control is implemented by either industry standards of controlled vocabularies such as SNOMED CT. There are very little enforcement mechanims and in many cases, industry standards are hotly contested

between different standards organisations or entities. This results in the least amount of effectiveness when it comes to authority control.

Q7.

The Best Thing and The Worst Thing

Explain both parts of this proverb from a modern philosopher of information organization:.

- *"the ease with which anyone can create a new vocabulary is the best thing about XML* **(5 points)**
- *… and also the worst thing."* **(5 points)**

A.

1. Though not as simple to grasp as HTML, XML is still a relatively easy way to separate a document's content from its presentation. Anyone can create a schema language for any conceivable purpose. The X means "extensible," and that's one of XML's great charms: it's infinitely extensible to any field that decides it wants or needs an XML language for whatever purpose. And since XML is "smart," creating an XML vocabulary almost automatically opens up a potential world of communication with other users (be they people or machines) in a way that plain text or even a markup language like HTML simply can't. And because XML can be validated against a document that defines what type of data can go in what field, it can be easy to see when something is wildly incompatible: if a field only takes an integer, but you've tried to put in a number, the validator will catch your mistake.

2. But should we all really be creating our own vocabularies? Is using a bad vocabulary really that much better than having no vocabulary at all? The same issues crop up when creating a schema language that we've seen in defining metadata elements and categories: we all see things differently, and language is an infinitely slippery thing, so even something as seemingly simple as OrderDate could mean different things to different people. Furthermore, there are a number of tradeoffs that go into creating a vocabulary, with decisions to be made about what features go into a model and the usability of that model once defined, and making those decisions takes time (and, in a business case, money). XML isn't self-describing, so just tossing up some element or attribute names doesn't necessarily equate to clear, reusable, clean content. Also, though I said an XML vocabulary can open up potential worlds of communication, the emphasis really should be on *potential*. Merely writing an XML vocabulary and a schema to define it doesn't mean your schema will be interoperable with anyone else's. Some sort of hub or neutral language may still be needed to create a common form for the schemas to interact — and creating that hub language raises the same set of issues (what elements to include? What to call them? What level?) all over again.

B.

"the ease with which anyone can create a new vocabulary is the best thing about XML **(5 points)**

- XML makes the creation of new data models very easy – there is no need to understand any specific language. Only knowledge of a domain is required
- The flexibility XML provides allows the creation of complex models representing varied domains
- Within the ecosystem of XML technologies, one can build a model, enforce and validate its use and selectively present part of the information represented by the model
- XML is also very expressive in comparison to the alternatives – the free form nature of tag assignment that XML allows makes it easy to have a rich vocabulary for a given domain

... and also the worst thing." **(5 points)**

- Proliferation of vocabularies – as it is very easy to work with, many people are compelled to reinvent the wheel
- There can be a tendency to over engineer when XML is involved due to the flexibility it provides. SOAP is a classic example ○ Verbosity – The free form nature of XML can lead to very verbose models when simple ones would suffice
- The result of the ease of use of XML is that it is used in many applications when it is not a good fit. The clash of the model centric and syntax centric world views that Tim Bray alludes to is a good example of this.
- Though creating a new vocabulary using XML is easy, maintaining it and ensuring that it achieves widespread adoption is very hard

Q8.

The Wisdom of Svenonius

"It has never been easy to explain why colossal labor should be needed to organize information."

Why is it hard to design a set of elements for use as metadata or as a descriptive vocabulary? **(4 points)**

What is it hard to design a system of categories? **(4 points)**

Why is it hard to define the authoritative form of a name? **(2 points)**

A.

Why is it hard to design a set of elements for use as metadata or as a descriptive vocabulary? **(4 points)**
When designing metadata or a descriptive vocabulary, you must make choices as to what features you will capture in your metadata model, as well as the granularity of your descriptors (how specific are they, how many works will fall under a particular heading). It is inherently expensive to assign metadata, so you must pick and choose carefully to attempt to capture only the most useful pieces of inofrmation. There is a trade-off between intelligent categorization and cost.
Additionally, different users have different needs, and it is hard to satisfy the needs of all. Well structured metadata about a book's property might be very useful to a librarian, while a book seller might care only about the ISBN.

- What is it hard to design a system of categories? **(4 points)**
  Categories are formal, structured things with formal meaning, but real-world distinctions are blurry. In class, Dr. Glushko explained an instance where Wittgenstein was asked to teach children a "game," so he taught them how to gamble. While people are familiar with what "game" means, there are items in the class that are considered more or less appropriate, and more or less representative of the class. Wittgenstein argues that there are no features "necessary and sufficient" to describe all members of a class, and that the meaning of a word depends on the word's use. In this context, it is difficult to ascertain what features embody the "it-ness" of a category. Even assuming you are able to pick the right features to describe one category, you still have to figure out how to make a complete system that captures all objects that you want to classify. If you assume that no two objects are exactly identical, then all categorization must necessarilly be lossy and reductionist, and ultimately imprecise.
- Why is it hard to define the authoritative form of a name? **(2 points)**
  Names are messy. People use variants of their names (with or without middle

initial/name, Bob vs. Robert, etc.), people use psuedonyms, and people and places can change their names. Spellings of names may vary based on translation issues (Beijing vs. Peking), author ignorance, or general disagreements on the correct spelling. With all of these issues, it can often be difficult to assign an "authoritative form," and once you do, it can still be costly and difficult to associate all the variants with the authority.

B.

Why is it hard to design a set of elements for use as metadata or as a descriptive vocabulary? **(4 points)**
The vocabulary problem, homonymy and synonymy make it very difficult to design a set of elements (as prescribed above). People come up with different terms for the same thing, the words I used to describe a dinner party in the vocabulary assignment were different from how other students approached the same problem. Furthermore, the same word may have multiple meanings (the plane flew, and the horizontal plane) or different words might express the same concept (cat, kitty, feline, etc).

· What is it hard to design a system of categories? **(4 points)**

A system of categories relies on a vocabulary (to describe it), so all of the problems above are inherited. Furthermore, the system of categories must be relevant to the items that are 'binned' into each category (classification). However, the classical theory of categorization (each item is a member based on 'necessary and sufficient' properties) hits a wall when attempting to categorize most things. Games is a good example, there are no set list of properties that define a game (of which every game has every property). Instead, we rely on Wittenstein's idea that 'meaning is use'. This means that our system of categories must account for items that aren't necessarily equivalents, but we treat them as equivalents. Bob used the example, "things you would take from your house if it was burning down". When the items that are placed into a category are not literal equivalents, it hard to design a system that effectively captures this.

· Why is it hard to define the authoritative form of a name? **(2 points)**

Language and spelling as well as format (Chung, Michael or Michael Chung) make defining an authoritative name difficult. Identifying and mapping all the variants to the authoritative can be difficult (to impossible). Another key question problem: authoritative to WHO? One person may believe that Chung should be the authoritative name (me!), but this is equivalent to Chueng (visa vi SoundEx).