

ISKO Singapore: Follow Up Questions on the NASA Taxonomy Case Study

These were follow up questions asked by participants in the session Joseph presented at the ISKO Singapore meeting on 20 January 2017, attended by participants in Singapore, and remote participants from Kuala Lumpur and Hong Kong. Thanks to Joseph for his careful replies!

DESIGN

Q: There is no People facet – I’m surprised by this. Why is this?

A: Good question. We do often create a People facet, or add people to the Organization facet. As I recall, the names of people did not come as a key access point when we did the fact finding for the NASA project.

Q: Is the Audiences facet used to restrict access or does it predict which group may use an item?

A: In the NASA Taxonomy, the Audience facet is intended to identify for whom the content has been created. Restrictions are handled by the Access and Access Controls facets. (Please see slide 17 and 18 at http://www.taxonomystrategies.com/wp-content/uploads/2017/01/NASA_Taxonomy.pdf)

Q: Instead of organization reinventing the wheel in Taxonomy, are there any industry specific taxonomies already available? Where can we find such taxonomies?

A: Yes there are many pre-existing taxonomy resources – some free, open source, and others available for license or purchase. However, we have found that most pre-existing resources are best used as a source rather than a ready to use plug-in.

- Taxonomy Warehouse - <http://taxonomywarehouse.com/>
- Wand Taxonomies - <http://www.wandinc.com/taxonomies.aspx>
- BARTOC (Basel Register of Thesauri, Ontologies & Classifications) - <https://bartoc.org/>

Q: Slide 19 shows that about 2,000 preferred terms were being used to describe some 200,000 resources. Is this typical of the scalability of taxonomy to content that you can expect with a faceted taxonomy?

A: The way to think about granularity in a faceted taxonomy is not the total number of terms, but by the number of facets and the number of terms in each facet. E.g., If you have 4 facets with 10 terms in each facet. The total number of possible pre-coordinated strings is 10^4 or 10,000. The following table is a listing of the number of concepts in each facet of the NASA Taxonomy.

Facet	Terms
Access	3
Access Controls	15
Audiences	67
Business Purpose	99
Content Types	105

Facet	Terms
Instruments	57
Locations	111
Missions and Projects	482
NASA Subject Categories	88
NASA Workforce Competencies	366
Organizations	658
Work Breakdown Structure	104

The total number of possible combinations is: 210,658,832,675,881,000,000,000.

IMPACT

Q: It looks like NASA Taxonomy has been there for quite some time. Is it still in use?

A: The Taxonomy was last updated in 2007, although it was uploaded to github in 2015. As far as I know, the Taxonomy is no longer in use. Although, as I mentioned in my talk, you can still observe some aspects of the taxonomy even today on the NASA.gov website. If you were to look into the document management systems at NASA Centers such as the Jet Propulsion Laboratory (JPL) I think you would see more evidence of the persistent use of the Taxonomy for describing work product documentation.

Q: Has the NASA Taxonomy been adopted by other industries? Do you know of any specific examples of organisations that have adopted the NASA Taxonomy?

A: As I mentioned in my talk, this was an early Taxonomy Strategies project, and we have been doing faceted taxonomy projects for large organizations ever since that time. Most of our projects reflect some of the characteristics of the NASA Taxonomy. I refer to these as “universal” facets such as Audience, Content Type, Business Purpose (Function), Industry, Location, Organization and Subject. I think these have become de facto common taxonomy facets. In terms of the specific value vocabularies in each of the NASA facets, these really vary from project to project. But Audience, Content Type, Business Purpose (Function), Industry and Location are good sources that can readily be re-purposed for use in other projects.

Q: How is the impact of using the NASA Taxonomy being measured by NASA? Is there any specific measurement or quantification? What are the main methods to measure or assess the effectiveness of a taxonomy?

A: As mentioned in the talk, an important value of the NASA Taxonomy was to get the CIO’s focused on website and content description policy beyond records management. Slides 26 and 27 of the presentation list the NASA Taxonomy benefits, but these were not quantified. However, I have built cost-benefit models for other clients where we identified specific values. Following is a table from a model we built for the U.S. Environmental Protection Agency (EPA) in 2006.

Service efficiency increase	Cost per unit	Current amount	% Increase/Decrease	Annual Benefit
Reduce FOIA Requests	\$ 500	5,000	10%	\$ 250,000
Reduce Other Information Requests	\$ 50	50,000	50%	\$ 1,250,000
Reduce cost per UU (unique user)	\$ 0.13	10,000,000	8%	\$ 108,333
Reduce time/cost to build super topic website	\$ 100,000	5	50%	\$ 250,000
Increase no. of webpages with metatags	\$ 3	50,000	100%	\$ 142,000
Decrease time/cost-to-regulation cycle	\$ 150,000	100	10%	\$ 1,500,000
Decrease time/cost to obtain permit	\$ 1,500	100,000	5%	\$ 7,500,000
TOTAL				\$ 11,000,333

Note: FOIA – Freedom of Information Act.

Other methods of evaluating taxonomy effectiveness include:

- Operational prototypes
- Facilitated focus groups
- User surveys
- Query log analysis
- Category usage analysis

to observe and measure trends; and

- Closed card sorting using online tools like Optimal Sort
- Finding content using online tools like Treejack
- Content tagging

to observe and measure consistency.

Q: In terms of outcomes, what was the feedback from users, user departments and organisations on the usefulness and effectiveness of the taxonomy?

A: During my time on the project we were focused on outreach. I think we had the greatest impact at the JPL where we had a champion in Jayne Dutra who worked with multiple groups. The goal was to provide a lingua franca across multiple repositories, and I think this was accomplished at JPL. In 2008, we received a Team award from the JPL CIO. More widely, I think the Taxonomy was inspirational and became a model and resource for others, and was incorporated into the overall navigation approach on NASA public websites. The overall NASA CIO got us on the agenda of the agency-wide CIO Council which helped to provide visibility for the Taxonomy.

PROCEDURAL

Q: Who were the people that you talked to in order to collect data for the taxonomy development, and who were the people who were identified to maintain the new Taxonomy?

A: We talked to librarians, archivists, data managers, CIO's and a few engineers across most of the 10 NASA Centers and at NASA Headquarters. The Taxonomy Committee included librarians, archivists and data managers – with a representative from each of the Centers.

Q: Have there been any new revisions to the developed NASA Taxonomy? How frequently does it need to be revised and what is the trigger to the revision?

A: As mentioned above, the last revision was in 2007. During its active period, I believe there were revisions published quarterly. During this time, most revisions were driven by taxonomy development. However, longer term, revisions were to be driven by user requests. The Committee was to evaluate the costs and benefits to decide on and schedule change requests. These are all taxonomy governance best practices which are similar to data management governance procedures.

Q: How did content in the 5 portals in which the taxonomy was implemented acquire metadata? Was there retrospective tagging?

A: There were 4 collections in the operational prototype built using the Siderean Seamark platform. The largest collection from the NASA Technical report Server (NTRS) was already tagged with some metadata. Retrospective tagging was done using “rough bulk tagging” methods, where simple business rules or entity extraction methods were used to automate the process. There was some light weight quality assurance to handle tagging anomalies. Other inaccuracies were corrected when they were discovered.

Q: What search engine does NASA use?

A: The operational prototype used the Siderean search engine. This was either Apache Lucene open source or a Verity license. Today, you would use Apache Solr to achieve this type of user experience. I don't know what NASA is currently using for search. I suspect lots of different search technology on different applications.

BIG PICTURE

Comment: When we talk about our work with our peers we frequently use unique experiences with a WOW factor. Building NASA's taxonomy is one of those. I do the same thing. In the past, I gave talks on document management & building the HK International airport. I have never worked on another project that had the scope, time sensitivity, complexity and political/social implications. Joseph implied a few times something similar. I'm not sure but I suspect these unique projects are not the best projects and experiences to learn from. The day-to-day mundane projects and experiences may be more useful. Comment?

JB: NASA was a special project for us, but I would call it foundational rather than exceptional. We have actually applied lots of what we learned at NASA in our more routine projects, e.g., fact finding interviews, the use of universal facets, the governance model, building taxonomy project websites, etc.

Q: If Joseph wants to share it would be useful to have his reflections on the similarities of the projects his firm has worked on over the past 10 years. Which techniques does his firm always or almost always use? Can the work on one project be recycled into another project? When his firm starts a project does it review past projects critically and carefully for work which can be reused? From the client perspective how are they similar? Is it number of employees, number of locations, type of industry, business model or something else.

A: We've learned along the way that every project no matter how large or small is different and always requires customization. There is no truly out of the box taxonomy, but there are good best practices about how to build, validate and maintain a taxonomy. For example, we thought that Content Types would be something that would be more or less standard from one project to the next. But this has not proven to be true. Sometimes, the whole project revolves around deciding just on the set of Content Types. On the other hand maybe we should not be surprised about this, because Content Types are foundational for implementing content management.

I stand by my comment that the size of the taxonomy doesn't change that much from small to large enterprise. What changes is the nature of the relationships between and among concepts. In our experience the key driver in complexity is whether the taxonomy will be used for content management, or whether it will be used something larger like program management; or to support inferencing, e.g. where the user asks about a symptom and we need to return a solution such as medical specialist, type of expert, machine part, etc.

Joseph Busch January 2017