THOUGHTS ON THE LD4PE RESOURCE CATALOGING PROCESS

SEAN DOLAN (DECEMBER, 2016)

KEY QUESTION: When a user clicks on a resource, based on a competency tagged to it, will they be satisfied or disappointed by the actual coverage given that competency by the resource in question?

The Cold Start problem:

The LD4PE Project began with many unknowns. The pool of available resources was not predefined, so it took a while to get a feel for what was out there. We were hoping to find ready-to-use tutorials and "recipes". We did find *some*, but these turned out to be a minority of resources. This presented a question- do we only catalog resources which closely fit our preconceived idea of a wholly desirable resource, or do we expand it to also include "less-than-ideal" resources which, nonetheless, had potential value to a teacher or student of Linked Data? For example: tools that could be used to demonstrate concepts; blog pages on which authors outlined solutions to consuming or publishing Linked Data which they had worked out on their own; video presentations which presented motivation and use cases for Linked Data, academic papers outlining cutting edge approaches to unsolved LD problems; technical specifications for ontologies and applications .... In other words, some of these resources were not necessarily ready to use "as is". PDFs might contain slides which clearly formed an outline of a lesson, but to which an educator would need to add exposition for them to be usable in a classroom. In this regard, whether a user is satisfied by a resource discovered through the LD4PE Catalog depends on the user's expectations.

How much information about a subject does the resource need to include to be "about" that resource?

Obviously, a mere mention of a concept is not enough – the goal is to catalog resources, not to index them. However, it soon became apparent that there existed few resources that tackled just one topic in great depth. Most resources (except for many of the SPARQL-related resources) were not this focused. Instead, they took a wider view (in a surprising number of cases, an almost "holistic" view) of the subject of Linked Data. For example, many started out with an explanation of LD principles and/or a description of the RDF data model, taking up half (or more) of the length of the resource, before finally adding on additional information on whatever subject was "advertised" in the title and abstract. Perhaps these resources *would* have been stronger and more useful if they had been more focused, instead of trying to "explain it all". However, it seems clear that the authors felt the need to supply background information before tackling their topic of interest. In some cases, the coverage of LD principles and the RDF data model- even though these topics were not the author's primary concern when creating the resource- was more thorough and of higher quality than some of the resources that were created by other authors to cover only these topics.

A Moving Target

The most difficult challenge was attempting to tag resources with competencies from a Competency Index (CI) that was a work in progress. A significant number of relevant *competencies* were not added to the CI until the final draft of the CI was produced about six months from the project's end. Whole *Topics*, a broad layer of granularity, were still fluid for nearly one year into the project.

A few are still not well-defined – e.g., what is meant by “RDF Data Analytics” and how is it different from “Assessing RDF Data Quality”? What does “Non-RDF Linked Data” mean?

Over the course of the Project (two years), two re-alignments of the Competency Index to all previously cataloged resources occurred. The first, in December 2015, coincided with an initial alignment to the newly formed *Topical Index*. At this point, resources were cataloged with the attitude that it was better to err on the side of “completeness” – i.e., make sure that every relevant concept that received coverage more substantial than a mere mention was tagged to the resource. There rationale for this approach was the following: 1) It was certain that a second re-alignment would take place in the future because the CI was still a work-in-progress; 2) The Content Partners were asking for gap analyses to determine the true extent of coverage of Linked Data topics in the pool of discovered resources to determine what new resources should be created, 3) The CI Editorial Board was, at that time, using resource warrant to fill in gaps in the CI.

The second re-alignment, completed between October and December 2016, occurred after the CI had been “finalized” and feedback on the Explore site had been gathered. Part of this feedback was that resources had been over-tagged, regarding both Topics and Competencies. During this second re-alignment, a conscious effort was made to limit tagging to one or two topics and no more than five competencies per resource (unless the resource was exceptionally long or was obviously intentionally broad in its coverage). Currently, the estimate is that between 90 and 95% of resource descriptions now meet these guidelines.

Making the cuts

Even with a much more complete CI and far greater knowledge of the pool of resources, this was still not an easy task. Many resources legitimately cover several Topics and a dozen or more Competencies. However, it was possible in most cases to say that a resource was slightly more relevant to one topic than another and, in this case, the cataloging decision was often made based on how many resources already existed which covered the “less relevant” topic. As previously mentioned, many resources covered Linked Data Principles and the RDF Data model as background material before diving into another topic, so these two topics were often the ones left un-tagged. It is often worth noting that, in quite a few cases, completely different Topics were now tagged to a resource from those that were deemed appropriate in previous catalogings. This mainly reflects an improved organizational structure of the CI by the end of the project. Also, Topics became much more clearly defined by the Competencies and Benchmarks which fell under them as gaps were filled in.

Similar “cuts” had to be made when tagging Competencies. Another frequent decision that had to be made regarding Competencies was whether to save space by selecting a broader Competency, or to be more precise by selecting the specific, narrower Benchmarks that fall under that Competency. In addition to how many Competencies the resource already had tagged to it, other considerations included whether the Benchmark was so specific that many users would not know to look for it (i.e., discoverability of a relevant resource would be enhanced by tagging to the broader Competency). On the other hand, if the narrower Benchmark were too often rejected in favor of the broader Competency, the number of tagged resources displayed to the user on the Explore site’s CI browse page would hover near zero, and they might think no resources existed which addressed these concepts.

Support for Searching and Browsing Behavior

More advanced users who come to the LD4PE site already knowing the specific topics that they are interested in learning about or teaching should be able to search the Competency Index only for those topics. Some users may be interested not in searching but in browsing, and if a title or descriptive blurb catches their eye, may click on a link to explore a resource. This seems especially important for those who are just entering the tangled web of serializations, standards, terminology, and tools that comprise the current Linked Data landscape. To these novice users, the finer distinctions between SPARQL as a query language and SPARQL ("Querying RDF data") as an Update language ("Manipulating RDF data") may not be clear; similar case with "Identify in RDF" vs "Managing identifiers (URIs)" and "Designing RDF Vocabularies" vs "Reasoning over RDF". Hopefully, users will take the time to skim the entire CI and pay special attention to both Topic Cluster and Topic headings before diving straight into the competencies. These headings provide the context necessary to understand why competencies and benchmarks have been organized a certain way. If a user does not find what they are looking for in one Topic, it is likely that it is in one of the neighboring Topics within the same Topic Cluster. Returning to the "Key Question", I believe that resources *are* now tagged "tightly" enough that most users will not feel misled by the Topics, Competencies, and Keywords tagged to the resources. If they also take the time to skim the descriptions, rather than relying solely on the titles of resources, the odds that they will be satisfied will only increase.

Purpose of Keywords

"About" keywords – as the project advanced, it became clear that these would not be used for filtering search results. Rather, they serve to highlight for the user at a glance some interesting aspects of the resource. So, it seemed unnecessary to duplicate terms that were already in the description (as they would be caught by full-text search) or terms found within the competencies tagged to the resources. During the second re-alignment, the number of "About" keywords tagged to resources was dramatically reduced. Some users might question why a resource about OWL or RDF Schema, for example, might not include "OWL" or "RDF Schema" as a keyword. Again, this is likely because these words already appear in the title and/or description, and/or the competencies tagged to the resource.