Reviewers' comments:  
  
Note: While submitting the revised manuscript, please double check the author names provided in the submission so that authorship related changes are made in the revision stage. If your manuscript is accepted, any authorship change will involve approval from co-authors and respective editor handling the submission and this may cause a significant delay in publishing your manuscript.  
  
  
**Reviewer #1:** GENERAL COMMENTS  
  
The paper presents a method to automatically create a faceted taxonomy for supporting faceted search over Linked Data. The authors first describe how a set of triples can be transformed to a materialized faceted taxonomy (they interpret each RDF property as a facet and a create a taxonomy of the property values based on their occurrences in the triples).  Then they report experimental results regarding the efficiency of the proposed transformation and about the quality of the produced taxonomies.  
  
In general, the addressed topic is important and the direction is interesting. Although the transformation is not difficult, it is good that the paper reported various experimental results over real datasets. It is also positive that the paper discusses the limitations of the proposed approach (in terms of false positive/negatives as regards automatic taxonomy production).  
  
Related work:  
The authors should refer to their previous work [1] and describe in the introduction what is the difference with the current paper.  
  
[1] Nansu Zong, Dong-Hyuk Im, Sung-Kwon Yang, Hyun Namgoong, Hong-Gee Kim: Dynamic generation of concepts hierarchies for knowledge discovering in bio-medical linked data sets. ICUIMC 2012: 12:1-12:5  
  
The presentation could be improved in several parts (see the detailed comments). The authors could also mention at the concluding sections topics and directions that are worth further work or research.  
  
Some questions:  
  
I wonder how the proposed transformation would behave in case the dataset has more complex structure. For example consider a dataset that have complex structures and uses blank nodes,  e.g.  {(Maria, address, b1) (b1, city, Seoul)}.  
Could facets corresponding to property-paths be defined?  
  
How the authors plan to exploit the derived materialized faceted taxonomy, i.e. what system they plan to use for offering  guided navigation?  
  
Do they plan to evaluate the browseable structure with users?  
  
  
DETAILED COMMENTS  
  
Section 1  
Section 1. 2nd line: Lacks - lacks  
  
Section 2  
  
Probably this section should contain 2 subsections  Background (e.g. on Linked Data) and Related Work. Section 2.2  would be better to be placed in Section 3.  
  
Section 2.1 Sometimes it is desirable to be able to specify the number of clusters (i.e. K in K-means).  
  
End of paragraph starting with (a):  It would be better to explain the term "meaningless".  
  
  
Refer and comment:  
  
[1] Nansu Zong, Dong-Hyuk Im, Sung-Kwon Yang, Hyun Namgoong, Hong-Gee Kim: Dynamic generation of concepts hierarchies for knowledge discovering in bio-medical linked data sets. ICUIMC 2012: 12:1-12:5  
  
  
Section 2.2.  
  
Before introducing your notation, first define RDF and Linked Data (and related notions: the RDF subject/predicate/object). After that you can define  your terminology and notations and describe how they map to RDF. Otherwise what an object o is, can be easily be e misunderstood.  
  
Def 2: Not very clear.  
  
I guess that it would be beneficial to introduce the running example at this point for making clear the definition.  
  
Figure 1: increase the font size  
  
Def 4:  relations or relationships?  
  
  
Section 3.  
Title: Change the title to  a more informative one  
  
Def 5:  Contains a typo.  
  
Don't start a line with a coma ","   (check the entire paper).  
  
  
Def 6:  It sounds a bit strange to define the intention as the object set.  
  
An example (ideally over the running example) could make clear what you want to say.  
  
Formula 3:   also provide an example (currently it cannot be understood).  
  
Describe the consequences of Def 7. What if instead of =1 we had =2?  
  
Section 3.2  
  
Figure 2: increase the font size.  
  
Figure 3: increase the font size. Currently the figure is not intelligible.  
  
Before referring to Figure 3 it is beneficial to say a few words for describing the input dataset (the one in (a) of Figure 3). The labels should be described too.  
  
Steps (a), (b), (c) aren't equivalent to the following?:  
  
Pairs(pi) = { (s,o) | (s,pi,o) \in Dataset }  
  
  
Section 4.1  
  
1st line: "Our object" or "Our objective"?  
  
Formula 4: How you select the threshold phi?  
  
  
Section 5.  
Update the introduction of section 5. You mention only performance there.  
  
Section 5.2.1  
Please explain how you treat the instances that do not have a value for a property.  
  
  
Section 6.1  
Also mention percentages not only ms (e.g. x %  less time).  
  
Section 6.2  
The times are in ms.  
Are all datasets loaded in main memory entirely? Make this clear.  
In general more details should be given regarding how these measurements have been performed.  
  
Figure 9: increase text font size  
  
Section 8  
In the concluding sections the authors could suggest methods or direction for tackling the issues described in Section 7?  
  
  
  
  
**Reviewer #2:** \* Text needs review of English. Sometimes it is not clear what the authors want to say.  
        ∘ Abstract:  
                ‣ "these data" --> "this data"  
        ∘ Introduction:  
                ‣ First sentence: Not 100% sure what you mean here  
                ‣ First paragraph: Rest is easier to understand, but not grammatically correct  
        ∘ "we used an open source called Colibri" --> "we used an open source project called Colibri"  
        ∘ Try to use shorter sentences!  
  
Introduction:  
\* the term "concept" and "instance" should be defined as early as possible  
\* Text: 1) for the first task, the proposed method takes 49 and 11,790 ms  
        ∘ Question: There is a big difference between 49 and 11,790 ms. What is the connection between these two numbers? Why is there a difference of over 11 seconds! --> Having read the rest of the paper, I see that the 49 ms is related to DBpedia and the 11 seconds to YAGO2. Please use this information here to help the reader to understand your text.  
\* Before referring to 0.917 and 0.780 --> specify that you mean F-measure/F-score. If it is F-Score, why is it greater than 1?  
\* Second mentioning of time (2,032 and 2,525 ms) --> Should may be written as: "our method takes between 2,031 and 2,515 ms"  
\*  Text: "1.65 for  Diseasome and  1.03 for  DrugBank"  
        ∘ Question: Are the numbers time in ms or F-Score?  
  
Related work:  
\* " Nature Language Processing" --> Should be Natural Language Processing  
\* When you refer to "single term" and "set of terms" it might be better to add additional information for readers unfamiliar with NLP methods. In addition it might be nice to mention how concept properties/features are extracted from the text. This is especially important, since Objects/Properties become a crucial part of the Facet Taxonomy  
  
3.2 Framework:  
\* Figure 3  
        ∘ How can you be sure that the two instances are really the same?  
        ∘ Why is DB00036 kept and DB00682 is filtered? There seem to be other diseases which have got more than one property (e.g. Disease 2949)  
        ∘ In my opinion the labeling of a to e differs from the labeling in figure 3. In my opinion the labelling of a.) should be placed after the current paragraph a.)  
  
4.2  
\* If we have Rule 1, what is Rule 2?  
\* Where were the filtered objects and instances stored? How do we get them back?  
  
5.  
\* What Java was it? Oracle, IBM?  
\* Maybe state the source ontologies which are used in the experiment together with OS and Java information.  
\* Do you have got any explanation why the number of unique instances is higher in YAGO2 than in DBpedia?  
\* How can you get a drop of more than 3 million instances for dbpedia down to only 348? Can you give a few examples?  
  
5.1.3;  
\* What is the difference between "efficiency and effectiveness" ?  
\* Why is "efficiency" not described in this section?  
  
6.1  
\* It might be nice to link "Definition 1" (and others) with the original definition.  
  
6.2  
\* Figure 9  
        ∘ It is not always possible to figure out what part of the chart belongs to what concept

====================================================================

**2nd time comments**

Reviewers' comments:  
  
  
**Reviewer #1:** I appreciate the efforts of the authors to tackle the comments of my review.   
The revised manuscript is much better than the previous one.  
  
Some comments about the revised manuscript follow.  
  
Section 1. Introduction  
"(1) for the first task, the proposed method takes 49 ms".   
More details should be given about the particular task that requires only 49 ms in the sense that to load DBpedia requires a lot of time. Since such absolute numbers depend on the machine, probably it is better in the introductory section to write something like: … it takes seconds (after having loaded them in main memory).  
  
  
I think it is worth mentioning in the introduction the weak points and the limitations of approach.  
  
Section 2  
  
Explaining the definitions through the running example has improved the readability of the paper. However some aspects are not very clear. One method to tackle this issue is to include in Section 2 also the input RDF triples (e.g. in turtle format), e.g. just before Figure 1.   
  
  
Section 3  
  
Figure 3 is still not very readable. Consider using a single column format instead of two column for that figure.  
  
Formula (3) is a bit problematic. The set over which c takes values should be included. Otherwise also the emptyset satisfies the condition.  
  
  
Section 4 and 6  
Any guideline on how one could select the value for f?  
This is important for cases where there is not any gold standard.  
  
  
Discussion.  
Other limitations or possible extensions (e.g. related to bnodes, or property paths) could be mentioned.  
  
For all sections:  
The use of language needs improvement.  
  
  
  
  
**Reviewer #4:** The paper tackles an important problem in LD: how to create a schema to the data taking into account the speed to which they grow. According to my point of view, besides the proposal itself, the main virtues of the paper are (1) the clarity with which it is written, and (2) the way in which the experiments have been documented. However, a deeper formal analysis of the proposal should be provided.  
  
The paper is biased to the empirical side of computer science. Nevertheless, a formal analysis, for example, of the complexity of the process, would give light to explain the results and the applicability conditions of the model.  
  
The omission of the formal side of computer science in this paper leads to the lack of explanation of the causes of the results of the experiments. For example, it is not clear for me why ICT is more efficient than FCA and subsumption.  
  
I also miss a section, or, at least, some paragraphs devoted to future directions. For example, it would be interesting explore the posibility of providing a systematic way to assign a priority to object properties to build sub-hierarchies.  
  
Finally, although figure 3 is very well designed to aid in the understanding of the paper, it is difficult to read the words that appear in it (they are too small).