Reviewer A:  
Review of “Encoding Provenance of Social Science Data: Integrating PROV  
with DDI”  
   
Overall, this is an interesting paper which should be accepted for the  
conference proceedings. There are, however a few issues which should be  
addressed, some minor and a few, like the direction of  
<prov:actedOnBehalfOf> which appear to be errors.  
   
A few suggestions are also listed at the end of the review    
Should page 2 paragraph 2, line 2 “created with King”  be “created  
what King”

Fixed

page 2 paragraph 2, line 6  extraneous “.”, sentence starts .A

Fixed

Figure 1 is very difficult to read when printed. It should be broken up with  
the diagram at top in vector format (preferably) or if not possible, at  
higher resolution. The text at the bottom should be text, not a bitmap.  
Figure 2 should be in vector format (preferably) or if not possible, at  
higher resolution.

Fixed

Figures 3, 4, 5 and 6 should be text, not bitmaps.

Fixed

Page 6 has typos in   “This is exactly our goal, to express within the  
DDI for \_\_\_ specific data set only its provenance dependencies and  
independently allow data sets to then express derivation from that existing  
data set \_\_\_ fire \_\_\_their own provenance bundle. ”

Fixed

While the text at the end of page 7 indicates that the XML on the pages  
following does not include a number of details in Figure 1, in Figure 4 some  
of the elements also do not seem to correspond to the graph in Figure 1.  
This might lead to some reader confusion. One example: the following does  
not seem to be in the graph.  
<prov:actedOnBehalfOf>  
<prov:delegate prov:ref=”cdr:USCB”/>  
<prov:responsible prov:ref=”cdr:ESMPD”/>  
<prov:activity prov:ref=”cdr:maintainElectronicVersion”/>  
</prov:actedOnBehalfOf>  
   
If might be best to have the XML correspond to features that are in Figure  
1.

Fixed

   
   
   
In section 3 should “•      UCSB: United States Census Bureau” be  
“•             USCB: United States Census Bureau”?

Fixed

Section 3.2  - the direction of ActedOnBehalfOf does not seem to agree with  
the XML (shown above) “the Economic and Statistical Methods Programming  
Division (ESMPD maintains the electronic version on behalf of the US Census  
Bureau (USCB)”  (prov:delegate acts on behalf of prov:responsible)

Fixed

   
Section 3.3 has the same issue with CES acting on behalf of the Census  
Bureau, but the XML indicating the opposite.  Census Bureau is misspelled  
as Census Euro (or is this suggesting the outcome of a U.S. debt default?  
;>) Also in 3.3 the text refers to porLBDPlan, but the XML refers to procLBD  
and procLBDplan

Fixed

   
There is an incomplete citation for Paul N. Edwards  is this it?:  
Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C.  
L., Ribes, D., Burton, M., & Calvert, S. (2013) Knowledge Infrastructures:  
Intellectual Frameworks and Research Challenges . Ann Arbor: Deep Blue.  
<http://hdl.handle.net/2027.42/97552>

Fixed   
   
   
Other suggestions:  
The last paragraph in section 4 could use a little more narrative about the  
graph in figure 7 and description of thoughts about the potential user  
interface. What is highlighted in the graph? How might a user use the  
visualization?

We have decided to move this figure since our work in this area is still so preliminary. We will focus on visualization in a future paper.  
   
In section 3 the paper describes the decision to use the DDI-Codebook  
framework due to the DDI-centric approach to provenance. It might be  
informative to discuss what might be lost by not integrating PROV with some  
profile of DDI-Lifecycle. Are there, for example, issues of loss of reuse of  
elements by reference?  
   
.  
We have inserted text in the paper that acknowledges that our choice of Codebook is one of expediency and that future work will focus on implementation within the context of Lifecycle. Also, as acknowledged in the paper, this will require some unraveling of the intersection of semantics between PROV and elements of Codebook.  
------------------------------------------------------  
  
------------------------------------------------------  
Reviewer B:  
  
The authors present an approach on including provenance represented in W3C's  
PROV model into DDI metadata. They describe how to connect PROV subgraphs to  
DDI metadata using the <relStudy> element and how to (re)use these subgraphs  
for several other DDI studies. This is supplemented by example XML code.  
This paper covers a relevant topic for researchers and institutes whose data  
is published in DDI, since provenance information, especially the origin of  
a data set and the processes it has been going through, is a highly valuable  
piece of information for researchers. The paper is clearly written and the  
examples are very helpful to illustrate the PROV subgraphs and their reuse.  
However, there are two minor issues that may enhance the paper.  
1. The actual connection between DDI and PROV with the <relStudy> element is  
described in the text and illustrates in Figure 2. Since there is already a  
lot of example XML code in the paper, it might be interesting to include  
also a few lines of DDI metadata including the <relStudy> element with a  
reference to the PROV subgraphs described in the following paragraphs. This  
would show how the PROV subgraph is actually addressed inside DDI and it  
would complete the line of examples presented in the paper.

The XML has been changed to show this inclusion.

2. The authors mentioned the ongoing work of a RDF representation of DDI  
metadata, which may easily support an accommodation of PROV subgraphs  
represented also in RDF. I think it may be helpful to delimit both  
approaches a bit clearer in the paper. I'm sure that there are application  
scenarios for both approaches (RDF and XML), since not everybody will  
publish DDI metadata in RDF. This may also strengthen the relevance of the  
presented work.

Language has been inserted in the introduction that hopefully addresses this point.

Overall, the paper presents an interesting and relevant approach for  
including provenance information to DDI metadata, which should be considered  
for a presentation at EDDI.  
  
------------------------------------------------------  
  
------------------------------------------------------  
Reviewer C:  
This is a clear discussion of PROV as used to relate provenance information  
(entity, agent, activity) at the level of a data set/study. Overall what  
seems to be missing from this paper is how and where you want this approach  
tied into DDI. My understanding of the level you are looking at addresses  
data capture from other existing sources (data source focus), the creation  
or use of synthetic data (process focus), and restructuring/reformatting of  
an existing data source. All of this at the level of an entity defined as a  
data source. I make that assumption based on your brief discussion of  
relStdy. As this is one of a series of papers I'm not sure if the source  
provenance at the cell level is covered somewhere else but it certainly is  
an issue

We agree that source provenance at the cell (variable) level is a potential issue, but it is a much more complicated issue, not least because the information about variable-level provenance is typically not available to third parties in the desired detail (an ongoing issue of replicability). Here, we tackle the (in real life) much easier situation of provenance of datasets first.

.Clarification Section 3 first paragraph:Rather than 2.X branch and  
3.X branch it is more correct to state that "DDI-Codebook, versions 1.0-2.X,  
primary focuses..." and "DDI-Lifecycle, beginning with version 3.0, is  
designed...." DDI-C encompasses 1.X and 2.X version strings and will  
continue to develop in the 2.X line. DDI-Lifecycle begins with version 3.0  
but will continue into additional major versions outside of the 3.X branch  
(4.X, 5.X etc). We trying to get people away from linking DDI-C and DDI-L to  
specific major version numbers as they both either currently or in the near  
future will encompass multiple major version numbers.

Agreed, we have formulated this section

"Some of the semantics  
of DDI-Lifecycle overlap and sometimes conflict with the PROV semantics  
specified by the W3C." This paper is for a DDI audience, you need to  
elaborate on this. "We argue that a DDI-centric approach to provenance, such  
as taken by the DDI-Lifecycle,..." Given that neither DDI-C or DDI-L has  
specifically addressed provenance (its only on the agenda) I would like to  
see more on what features currently in DDI (either model) can be used to  
address provenance, where they work and where they don't. I get a clear view  
of how PROV works, I know that it is mapped to PREMIS where DDI has looked  
in the past to identify what is needed to address questions of provenance at  
the study level and down to the data item/cell level. What I'm not seeing is  
a clear means of how the PROV approach could/should be integrated into DDI.  
You specify one DDI-C element <relStdy> but show nothing regarding how you  
would extend this element. We are all aware that DDI does not currently have  
either a good provenance model nor a good process model. I bring up process  
models as they have the same basic structure (entity, agent, activity plus a  
number of additional specifications) and I think we should be discussing  
these two perspectives together as DDI move forward. I think that a  
discussion of how and at what levels the PROV model could be integrated into  
DDI, used in conjunction with DDI, or mapped to DDI would be very helpful.

Hopefully, we have addressed this through some clarification in the text. Basically, we demonstrate one possibility of integration in the context of DDI-Codebook. We recognize that there are other methods. The key issue is to make these usable by both the data-creator and the archive-maintainer and easily interpretable by machines for visualizations, etc.

Otherwise this becomes a paper identifying what PROV contains and how it  
works. There is a bit in section 3 that notes "making minor change to the  
DDI 2.5 schema to instruct validation tools to evaluate the PROV sub-tree  
within the constraints of the PROV XML Schema but I don't get a sense of  
where you want this to be available. As for DDI-L we are working towards a  
UML model with multiple implementation languages (XML, RDC to start with)  
and much of the RDF structure at the moment is comprised of tags from other  
languages (OWL, SKOS, Dublin Core, etc.). Is it a viable approach to suggest  
the use of PROV tags and structure for items in and RDF expression of DDI-L.  
Clearly PROV provides structure and content needed in both development track  
of DDI, I'd just like to see more on how this structure and content can be  
integrated into DDI. I want to see more on the specific problems within  
DDI-C and DDI-L in terms of what it does and doesn't do, what it needs, and  
how PROV can solve the problem

we have modified the language in the paper clearly indicate that what we propose here, exclusively XML-based approach, is one alternative. We agree that it is a viable approach to incorporate PROV properties and classes in the RDF expression of DDI-Lifecycle. However, we believe that in the end both approaches, pure XML and pure RDF, should be defined and should be semantically equivalent, allowing implementers to choose the approach that they are most comfortable with. The UML modeling approach sounds like the right way to go. Hopefully our main example in the paper indicates why PROV semantics are at least necessary within the scope of DDI