A SAS Dataset Use Case For Enhanced Data Citation

Larry Hoyle, Institute for Policy & Social Research, University of Kansas

At Dagstuhl event 14432 a group funded by NSF grant 1448107 created several use cases, including one where a quantitative dataset would be created from a qualitative dataset through text mining. For the latter we chose the raw minutes from the first three days of the meeting. The derived dataset would have as its unit of analysis the topics computed by the text mining software. We decided to also create an example variable to show how it might have source information useful for a citation. It also became clear that the text mining procedure could itself serve as an example instrument, in that it is essentially a "black box" with a set of inputs – data and parameters, and an output – a dataset. A source information package for this procedure would include documentation of all of these inputs.

We took the minutes of the first three days of the workshop from three separate Google Docs, exported them into Microsoft Word and appended them into a single text file in Ultraedit. This process made each paragraph in the original documents into a single line in the text file.

The following SAS program read the minutes into a SAS dataset (Node "ReadRawMinutes" in Figure 3):

```
filename mins "C:\DDRIVE\projects\various\DDI\NSFDearColleague\MineMinutes\Oct20_22Minutes.txt";

libname minlib "C:\DDRIVE\projects\various\DDI\NSFDearColleague\MineMinutes";

data minlib.minutes;
infile mins lrecl=520 pad;
input para $520.;
run;
```

The SAS Enterprise Miner, Text Miner process shown in Figure 1 produced a Topics dataset and a Clusters dataset from the Minutes dataset using the default options. These options are listed in Appendix 2.

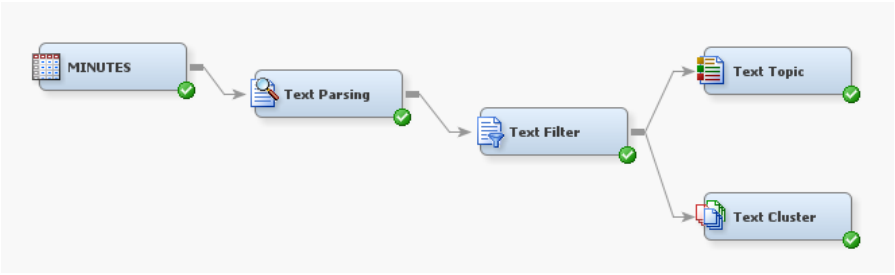

Figure 1 - Screenshot of the Text Miner Process

We saved the Topics results as a SAS Dataset from the Results Window for the Text Topics node.

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.402	0.287	+publisher,titl...	7	36
Multiple	2	0.389	0.280	+object,citabl...	5	40
Multiple	3	0.331	0.272	+element,+cit...	7	81
Multiple	4	0.327	0.271	+cite,+dataset...	14	56
Multiple	5	0.309	0.270	data,+citation,...	9	67
Multiple	6	0.347	0.269	metadata,+cit...	14	95
Multiple	7	0.310	0.265	ddi,nature,+rol...	11	75
Multiple	8	0.283	0.261	+contribution,...	10	38
Multiple	9	0.279	0.248	+resource,+k...	12	23
Multiple	10	0.272	0.255	+role,+contrib...	8	57
Multiple	11	0.253	0.256	datacite,meta...	14	75
Multiple	12	0.279	0.251	information,+c...	14	91
Multiple	13	0.244	0.254	+citation,differ...	20	89
Multiple	14	0.252	0.241	+question,+el...	11	50
Multiple	15	0.234	0.241	+author,+publi...	10	42
Multiple	16	0.231	0.240	+variable,+reu...	17	57

Figure 2 - The Topics Table in the Results Window

The resulting SAS dataset was then modified in SAS Enterprise Guide. A new variable was computed, combining the topic number, the number of documents using the topic and the list of most highly weighted terms for the topic. Since the variable names for the Topic Result table are standard, this is a reusable variable which can be recomputed from the variables (_n_, _NumDocs,_name) using the following transformation.

```
TopicDescription = catx(" ","Topic ",_n_," has ",_numDocs," documents and",_name," as Terms:");
```

We used the Topics2 dataset and the new variable (TopicDescription) as objects to be cited.

Additional metadata corresponding to DDI3.2 elements were added to the SAS dataset and a DDI3.2 instance and a codebook were generated from that (See Appendix 1 and Appendix 4). The complete set of extended attributes for the dataset is listed in Appendix 5. Extended attributes for the variable *TopicDescription* are listed in Appendix 6.

Figure 3 shows the Enterprise Guide process flow diagram. Text Miner is a separate application so that is represented in the flow by a note.

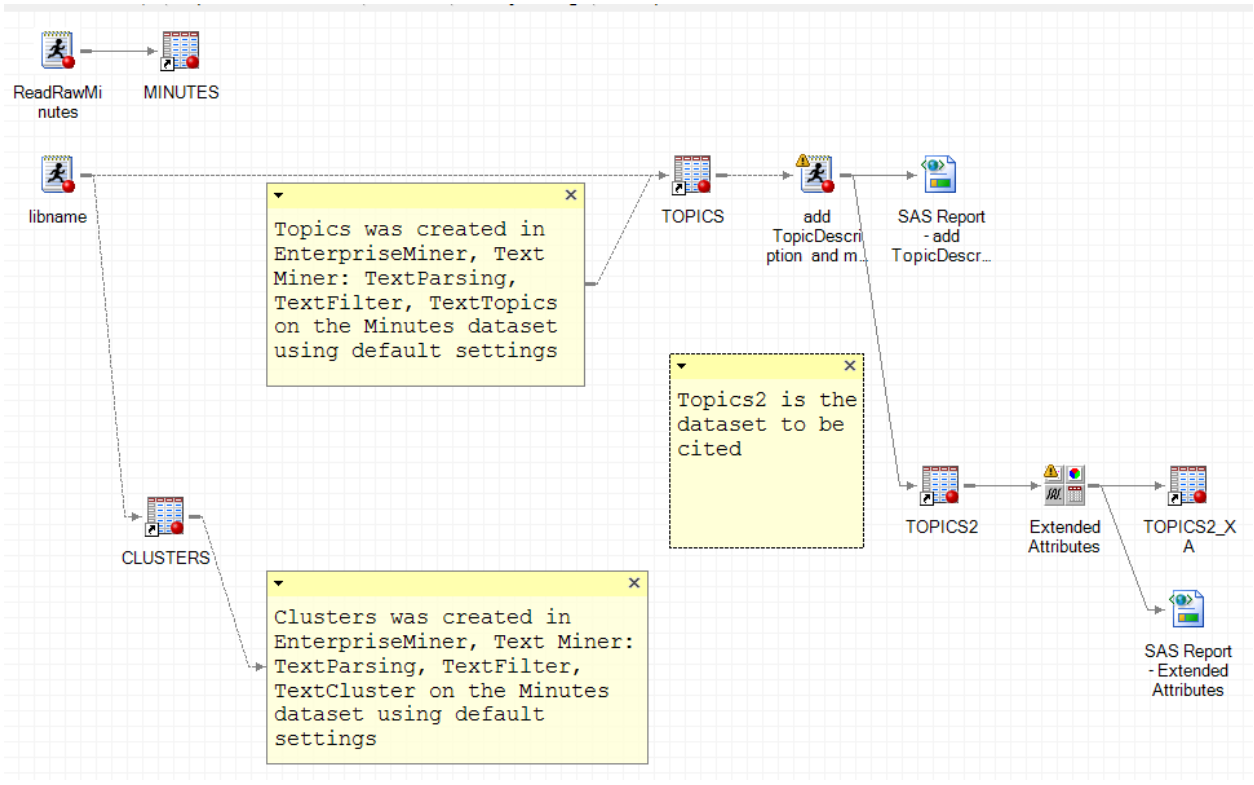

Figure 3 - Enterprise Guide Process Flow

In DDI3.2 the SAS Extended Attribute Name-value pairs can be represented as <r:UserAttributePair> elements. For example:

```
<r:UserAttributePair>
      <r:AttributeKey>Study_FundingInformation</r:AttributeKey>
      <r:AttributeValue>Participant travel and accommodations funded by NSF grant 1448107</r:AttributeValue>
 </r:UserAttributePair>
```

DDI 3.2 Citation elements can contain the following elements.

AlternateTitle, Contributor, Copyright, Creator, dc:any, InternationalIdentifier, Language, PublicationDate, Publisher, SubTitle, Title (Note that dc:any can be replaced by any Dublin Core element.)

Citation Information for the Dataset

The DDI3.2 Citation element for the sample dataset can read:

```
<r:Citation>
      <r:Title>
        <r:String>Topics generated from minutes from NSF1448107 group at Dagstuhl event 14432</r:String>
      </r:Title>
      <r:Creator>
        <r:CreatorName>
          <r:String>Larry Hoyle</r:String>
        </r:CreatorName>
      </r:Creator>
      <r:Publisher>
        <r:PublisherName>
          <r:String>University of Kansas</r:String>
        </r:PublisherName>
      </r:Publisher>
      <r:Contributor>
        <r:ContributorName>
          <r:String>Mary Vardigan(conceptualization, equal), Sam Hume(conceptualization, equal), Sanda Ionescu(conceptualization, equal), Jay Greenfield(conceptualization, equal), Jeremy Iverson(conceptualization, equal), John Kunze(conceptualization, equal), Barry Radler(conceptualization, equal), Stuart Weibel(conceptualization, equal), Michael C. Witt(conceptualization, equal)</r:String>
        </r:ContributorName>
      </r:Contributor>
      <r:PublicationDate><r:SimpleDate>2014-11-17</r:SimpleDate></r:PublicationDate>
      <dc2:description>This dataset is intended as an example for attaching source information to a dataset and a variable.  from SAS Dataset:
  name: TOPICS2
  label: A SAS Dataset generated from minutes from NSF1448107 group at Dagstuhl event 14432
 observations:        25
  variables:        8
  encoding: wlatin1  Western (Windows)</dc2:description>
      <dc:created>  2014-11-09T11:43:30.8</dc:created>
     <dc:modified>  2014-11-09T11:43:31.0</dc:modified>
      <dc2:language>en-US</dc2:language>

      <dc:abstract>This dataset was created from the October 20-22 minutes of the NSF1448107 sponsored group attending Dagstuhl event 14432. Topics were generated using default settings of SAS Text Miner on the  concatenated raw minutes files, one record per paragraph,  for the first three days of the meeting.</dc:abstract>
      <dc:accessRights>Freely available, with attribution</dc:accessRights>
      <dc2:contributor>Mary Vardigan(conceptualization, equal), Sam Hume(conceptualization, equal), Sanda Ionescu(conceptualization, equal), Jay Greenfield(conceptualization, equal), Jeremy Iverson(conceptualization, equal), John Kunze(conceptualization, equal), Barry Radler(conceptualization, equal), Stuart Weibel(conceptualization, equal), Michael C. Witt(conceptualization, equal)</dc2:contributor>
      <dc2:rights>Freely available, with attribution</dc2:rights>
      <dc:spatial>Schloss Dagstuhl, Wadern, Germany</dc:spatial>
```

```
  <dc:temporal>2014-10-20 to 2014-10-22</dc:temporal>
  <dc2:subject>Enhanced citation</dc2:subject>
</r:Citation>
```

All of the elements we propose for versionable objects appear here with the following caveats. Contributor is shown here as a structured string including both *role* and *degree of contribution*. DDI 3.2 would allow multiple <r:Contributor> elements, each of which could include a r:<ContributorRole>, but not a degree of contribution. Copyright information was not provided but could be structured in an <r:Copyright> element. License appears as <dc:accessRights>.

In many small research projects UserID may not be clear in the context of a dataset which is being developed. A dataset has a name, unique within a file system folder, but not necessarily unique outside of that context. A DDI identifier is not certain to remain the same during development of the file. Once archived, the dataset will probably have a unique identifier.

UserAttribute pairs are recorded separately within the DDI3.2 instance – on each element to which they apply. The user attributes for this dataset are listed in Appendix 5.

The citation above also includes coverage information – spatial, temporal, and topical (subject).

Note also that creation date, modification date and publication date are all listed.

Citation Information for a Variable

We created a new variable (TopicDescription) that could be reusable with the Topic Result table from any SAS Text Miner Text Topic node instance. The Citation information for that variable is different than for the dataset. DDI3.2 Doesn't allow an r:Citation element to be attached to a variable, so the citation information was structured in a set of r:UserAttributePair elements. These are listed in Appendix 6. Note that the Creator and Contributor information for the variable is different than for the dataset as a whole. Allowing source description information to be attached to any versionable object in DDI4 will make the structure of this information more consistent.

W5H SP

During the meeting, we discussed the requirements of source information for a dataset as including information about who, what, when, where, whether, how, structure, and provenance.

The DDI 3.2 instance we generated for our example dataset includes

Who – Creator, Contributor, FundingInformation
We did not include institutional responsibility or a reference to a persistent researcher identifier. DDI3.2 allows both Creator and Contributor to reference a structure which can contain references too external persistent identifiers (like ORCID). It is not clear how to document institutional responsibility for different phases of the data lifecycle in DDI3.2.

What - Title, Description, Abstract, Version

When - Creation Date, Modification Date, and Publication Date
In some datasets there may be other date/time references required. Retrospective studies may ask respondents about some time period in the past. These can be documented in r:TemporalCoverage

Where - Publisher, Pointer to metadata, Actionable link to dataset
We provided a Handle pointing to a landing page having links to the data files and associated metadata. This page is not really structured to provide an actionable link to each dataset.

Whether - Access Rights, Copyright, License, Permanence
It is not clear whether AccessRights and License are both required.

How - ProcessingDescription, GenerationInstruction, Language, CollectionMethodology, RelatedResource…
The metadata includes descriptive text about the method used to generate the dataset, its source, collection methodology.

Structure - LogicalProduct, PhysicalInstance …
A full set of DDI3.2 description of the data could be harvested from the SAS dataset. This should allow machine actionable interpretation of the structure of the dataset.

Provenance – Only as unstructured descriptions in CollectionMethodology, ProcessingDescription, and GenerationInstruction.

Instrument

The Text Mining procedure used to produce the Topics dataset can be considered as "black box" instrument that takes a text dataset and produces a quantitative dataset. Documenting the use of this instrument to allow someone to reproduce the results requires recording all of the parameter choices made in using this instrument. This instrument description description-type can require a large number of information objects unique to the particular instrument. Appendix 2 shows the values of the 96 properties set for the run of Text Miner used to generate the topics dataset. At the time of this analysis these were the "default" choices, but there is no guarantee that the default values will remain the same for future versions of the software so listing them is important for replication. SAS Enterprise Miner allows the export of diagram properties as an XML file. The tables shown here were processed from that file.

In the case of Text Miner many properties are relevant to only one node in the process. This can be seen with the properties *delimit, bCapitalize, bPartOfSpeech, NounGroups, multiDS, bPatterns, stopList, ignorePOS, ignoreAttrib, bStems, synonymDS*, shown in Table 1, which are relevant only to the Text Parsing node. Each of the nodes in this process has its own set of inputs and outputs and might be considered sub-instruments linked by their inputs and outputs.

Property	Node_ TextParsing	Node_ TextFilter	Node_ TextTopic	Node_ TextCluster
delimit	Std			
bCapitalize	Y			
bPartOfSpeech	Y			
NounGroups	Y			
multiDS	SASHELP.ENG_MULTI			
bPatterns	NONE			
stopList	SASHELP.ENGSTOP			
ignorePOS	'AUX' 'CONJ' 'DET' 'INTERJ' 'PART' 'PREP' 'PRON'			
ignoreAttrib	'NUM' 'PUNCT'			
bStems	Y			
synonymDS	SASHELP.ENGSYNMS			

Table 1 – Some parameters Relevant Only to Text Parsing

Some of these properties, like *stoplist* point to a data file (SASHELP.ENGSTOP). Appendix 3 shows the contents of the *Obs_TERM* variable from the stoplist dataset. This is a list of terms that will be ignored in the computations within and following the text parsing node. In this case, then, an input parameter can be complex – e.g. the contents of another dataset.

Table 2 shows the connections among nodes as a "from-to" dataset. This table allows reproduction of the process flow diagram (in Figure 1).

CONNECTION_FROM	CONNECTION_TO
Ids	TextParsing
TextParsing	TextFilter
TextFilter	TextTopic
TextFilter	TextCluster

Table 2 – Connections Among Nodes

Sample Citations

Here we show how citations for the dataset and the variable might be listed in three common styles.

Dataset

APA – Hoyle, Larry (2014). *Topics generated from minutes from NSF1448107 group at Dagstuhl event 14432* [data file, codebook, DDI metadata] http://hdl.handle.net/1808/15746.

MLA - Hoyle, Larry. *Topics generated from minutes from NSF1448107 group at Dagstuhl event 14432.* University of Kansas, 2014. Web. 17 Nov 2014.

Chicago - Hoyle, Larry. *Topics generated from minutes from NSF1448107 group at Dagstuhl event 14432.* Lawrence Kansas: University of Kansas. 2014. http://hdl.handle.net/1808/15746.

All three styles leave out contributors:

Contributors: Mary Vardigan(conceptualization, equal), Sam Hume(conceptualization, equal), Sanda Ionescu(conceptualization, equal), Jay Greenfield(conceptualization, equal), Jeremy Iverson(conceptualization, equal), John Kunze(conceptualization, equal), Barry Radler(conceptualization, equal), Stuart Weibel(conceptualization, equal), Michael C. Witt(conceptualization, equal)

Variable - TopicDescription

APA – Hoyle, Larry (2014). *Topic Descriptor Combining Sequence Number, Number of Related Documents, and Terms List From A SAS Text Miner Text Topics Node Result Table*[variable]. http://hdl.handle.net/1808/15746.

MLA - Hoyle, Larry. *Topic Descriptor Combining Sequence Number, Number of Related Documents, and Terms List From A SAS Text Miner Text Topics Node Result Table*. University of Kansas, 2014. Web. 17 Nov 2014.

Chicago - *Topic Descriptor Combining Sequence Number, Number of Related Documents, and Terms List From A SAS Text Miner Text Topics Node Result Table*. Lawrence Kansas: University of Kansas. 2014. http://hdl.handle.net/1808/15746.

Contributor: Mary Vardigan(writing – review & editing, lead)

In the three examples above only the APA style indicates that the object being cited is a variable. The MLA style doesn't yield a persistent identifier. The handle shown above points to a landing page with a description of the collection and more than a dozen URLs to objects within the collection (original raw data, software code, a codebook, a DDI instance). An explicit link to the data file and another explicit link to the structured metadata for the data would be much more machine actionable.

None of these styles allow for designation of a role or degree of contribution for the creator or a listing of contributors and their roles. If the standard citation styles included an explicit reference to structured metadata, including some mechanism for identifying the structure style, both of these problems could be handled by machine actionable searching of the metadata.

Implications for DDI4

DDI4 will need a mechanism to allow the incorporation of a set of information objects with a vocabulary drawn from an external controlled vocabulary. Ideally this mechanism will include the capability of validation those objects and also indicating relationships among the objects. Designing this mechanism will be a task for the DDI4 modeling group.

This need comes up both for instrument parameters and for the vocabulary for Creator and Contributor roles. The latter might have a hierarchical structure. DDI3.2 allows for attaching role to Contributor but not Creator. In a large study co-principal investigators may have specialized roles which should be documented. In each case role should also be paired with a "degree of Contribution" measure. We propose using the CRediT taxonomy (Allen et al) as the top level of a taxonomy for describing role and a three level category (e.g. "lead", "peer", and "supporting") for degree as in the CRediT proposed standard.

Input parameters may also be complex objects, including datasets, as noted for the stoplist dataset. Parameters might also come from stream sources at specific times.

The group recommendation to allow source description information to be attached to any versionable object will yield a more consistent structure for this information. For citation type information the addition of Role and DegreeOfContribution to Creator and Contributor, along with the elements already present in DDI3.2 should allow for a usable set of information.

For other source information types though, DDI4 will need to support external controlled vocabularies for attribute names and complex data types (including datasets) for attribute properties.

References

Allen, Liz, Jo Scott, Amy Brand, Marjorie Hlava & Micah Altman (2014). *Publishing: Credit where credit is due?* Nature 508, 312–313 (17 April 2014), doi:10.1038/508312a.

Appendix 1 – SAS Program Creating Variable TopicsDescription and Adding Metadata

```
data CITE.TOPICS2(label='A SAS Dataset generated from minutes from NSF1448107 group at Dagstuhl
event 14432');
set CITE.TOPICS;
Length TopicDescription $ 1000;
TopicDescription = catx(" ","Topic ",_topicID," has ",_numDocs," documents and",_name," as
Terms:");
 run;


title  'Contents of the Revised Dataset';proc datasets lib=CITE nolist ;
     modify TOPICS2 ;
        xattr options seglen = 4000;
            XATTR SET DS Abstract='This dataset was created from the October 20-22 minutes of the
NSF1448107 sponsored group attending Dagstuhl event 14432. Topics were generated using default
settings of SAS Text Miner on the  concatenated raw minutes files, one record per paragraph,  for
the first three days of the meeting.' ;
            XATTR SET DS AccessRights='Freely available, with attribution' ;
            XATTR SET DS Contributor='Mary Vardigan(conceptualization, equal), Sam
Hume(conceptualization, equal), Sanda Ionescu(conceptualization, equal), Jay
Greenfield(conceptualization, equal), Jeremy Iverson(conceptualization, equal), John
Kunze(conceptualization, equal), Barry Radler(conceptualization, equal), Stuart
Weibel(conceptualization, equal), Michael C. Witt(conceptualization, equal)' ;
            XATTR SET DS Creator='Larry Hoyle' ;
            XATTR SET DS Description='This dataset is intended as an example for attaching source
information to a dataset and a variable.' ;
            XATTR SET DS FundingInformation='This dataset was created during Dagstuhl event 14432
by a group funded from NSF grant number1448107.' ;
            XATTR SET DS Language='en-US';
            XATTR SET DS License='Freely available, with attribution' ;
            XATTR SET DS ParentDatasets='Cite.Minutes, Cite.Topics' ;
            XATTR SET DS Permanence='Permanent: Unchanging Content' ;
            XATTR SET DS PublicationDate='2014-11-17' ;
            XATTR SET DS Publisher='University of Kansas' ;
            XATTR SET DS RelatedResourceAuthor='Jay Greenfield, Larry Hoyle, Sam Hume, Sanda
Ionescu,  Jeremy Iverson, John Kunze, Barry Radler, Mary Vardigan, Stuart Weibel, Michael C. Witt'
;
            XATTR SET DS RelatedResourcePublicationDate='2014-11-17' ;
            XATTR SET DS RelatedResourcePublisher='University of Kansas' ;
            XATTR SET DS RelatedResourceRelationship='isDerivedFrom' ;
            XATTR SET DS RelatedResourceTitle='Minutes for Oct 20-22 2014 from NSF1448107 group at
Dagstuhl event 14432' ;
            XATTR SET DS ResourceType='dataset' ;
            XATTR SET DS SpatialCoverage='Schloss Dagstuhl, Wadern, Germany' ;
            XATTR SET DS Study_AnalysisUnit='paragraphs from raw minutes files' ;
            XATTR SET DS Study_CollectionMethodology='Minutes were generated as Google Docs, one
for each day at the Dagstuhl workshop. All participants could simultaneously edit the daily
minutes file. Minutes for 2014-10-20, 2014-10-21, and 2014-10-22were copied from  downloaded
Microsoft Word files and concatenated into a single text file. This file was read into SAS and
then used as input for SAS Text Miner with all default options chosen. The Topics results table
was exported as this SAS dataset' ;
            XATTR SET DS Study_FundingInformation='Participant travel and accomodations funded by
NSF grant 1448107' ;
            XATTR SET DS Study_KindOfData='SAS Text Miner Topics results table. Derived topics
descriptions, dataset includes metadata in SAS extended attributes' ;
            XATTR SET DS Study_ProcessingDescription='/* data read from concatenated minutes */
filename mins "C:\DDRIVE\projects\various\DDI\NSFDearColleague\MineMinutes\Oct20_22Minutes.txt";
libname minlib "C:\DDRIVE\projects\various\DDI\NSFDearColleague\MineMinutes";    data
minlib.minutes;  infile mins lrecl=520 pad;  input para $520.;  run; /* Dataset CITE.Topics
generated from SAS Text Miner from minlib.minutes,  CIte.Topics2 then generated by */  data
CITE.TOPICS2;   set CITE.TOPICS;   Length TopicDescription $ 1000;  TopicDescription = catx("
","Topic ",_topicID," has ",_numDocs," documents and",_name," as Terms:");   run;' ;
            XATTR SET DS Study_Purpose='a sample dataset for enhanced data citation' ;
            XATTR SET DS TemporalCoverage='2014-10-20 to 2014-10-22' ;
            XATTR SET DS Title='Topics generated from minutes from NSF1448107 group at Dagstuhl
event 14432' ;
            XATTR SET DS TopicalCoverage='Enhanced citation' ;
            XATTR SET DS Version='1.0' ;
```

```
            XATTR SET DS VersionDate='2014-11-17' ;
            XATTR SET DS VersionResponsibility='Larry Hoyle' ;
            XATTR SET VAR TopicDescription (AccessRights='Freely available, with attribution') ;
            XATTR SET VAR TopicDescription (AnalysisUnit='paragraphs') ;
            XATTR SET VAR TopicDescription (Concept='A label for a topic generated by SAS Text
Miner combining the topic number, the number of documents relating to teh topic and the key
descriptive terms for the document.') ;
            XATTR SET VAR TopicDescription (Contributor='Mary Vardigan(writing – review & editing,
lead)') ;
            XATTR SET VAR TopicDescription (Creator='Larry Hoyle') ;
            XATTR SET VAR TopicDescription (Description='A variable to be used with a Topics
results dataset produced by SAS Enterprise Miner Test Miner. One string includes topic number,
number of related Documents, and key terms.') ;
            XATTR SET VAR TopicDescription (GenerationInstruction='TopicDescription = catx("
","Topic ",_n_," has ",_numDocs," documents and",_name," as Terms:");') ;
            XATTR SET VAR TopicDescription (LevelOfMeasurement='Nominal') ;
            XATTR SET VAR TopicDescription (Language='en-US') ;
            XATTR SET VAR TopicDescription (Permanence='Permanent: Unchanging Content') ;
            XATTR SET VAR TopicDescription (ProcessingDescription='computed with the following SAS
assignment statment: TopicDescription = catx(" ","Topic ",_topicID," has ",_numDocs," documents
and",_name," as Terms:"); _topicID, _numDocs, and _name are standard variable names from an
unsorted Topics dataset saved from Enterprise Miner.') ;
            XATTR SET VAR TopicDescription (PublicationDate='2014-11-14') ;
            XATTR SET VAR TopicDescription (Publisher='University of Kansas') ;
            XATTR SET VAR TopicDescription (ResourceType='Variable') ;
            XATTR SET VAR TopicDescription (Role='Potentially useful for topic labeling') ;
        XATTR SET VAR TopicDescription (Title='Topic Descriptor Combining Sequence Number, Number
of Related Documents, and Terms List From A SAS Text Miner Text Topics Node Result Table');
            XATTR SET VAR TopicDescription (VariableIdentifier='TopicDescription') ;
            XATTR SET VAR TopicDescription (Version='1.0') ;
            XATTR SET VAR TopicDescription (VersionDate='2014_10_23') ;
            XATTR SET VAR TopicDescription (VersionResponsibility='Larry Hoyle') ;
     contents data = TOPICS2 ;run; quit;
```

Appendix 2 - Selected Properties for SAS Enterprise Miner Text Miner Generation of Topics

Property	Node_DataSource	Node_TextParsing	Node_TextFilter	Node_TextTopic	Node_TextCluster
DataSource	minutes				
Scope	LOCAL				
Role	RAW				
Library	MINLIB				
Table	MINUTES				
NCols	1	.	.	.	.
NObs	867	.	.	.	.
NBytes	459776	.	.	.	.
Segment					
OutputType	VIEW				
ForceRun	N	N	N	N	N
ComputeStatistics	N				
DataSelection	DATASOURCE				
NewTable					
DataSourceRole	RAW				
MetaAdvisor	BASIC				
ApplyIntervalLevelLow	Y				
IntervalLowerLimit	20	.	.	.	.
ApplyMaxPercentMissin	Y				
MaxPercentMissing	50	.	.	.	.
ApplyMaxClassLevels	Y				
MaxClassLevels	20	.	.	.	.
IdentifyEmptyColumns	Y				
VariableValidation	STRICT				
NewVariableRole	REJECT				
DropMapVariables	Y				
DsId	minutes				
DsSampleName					
DsSampleSizeType					
DsSampleSize					
DsCreatedBy	lhoyle				
DsCreateDate	1729690204.5	.	.	.	.
DsModifiedBy	lhoyle				
DsModifyDate	1729690204.5	.	.	.	.
DsScope	LOCAL				
Sample	D				
SampleSizeType	PERCENT				
SampleSizePercent	20	.	.	.	.
SampleSizeObs	10000	.	.	.	.
DBPassThrough	Y				
RunAction	Train	Train	Train	Train	Train

Description					
Location		CATALOG	CATALOG	CATALOG	CATALOG
Catalog		SASHELP.EMTXTEXT.PARSE.SOURCE	SASHELP.EMTXTEXT.FILTER.SOURCE	SASHELP.EMTXTEXT.TOPIC.SOURCE	SASHELP.EMTXTEXT.CLUSTER.SOURCE
resolution					LOW
maxK	.	.	.	.	100
exactOrMaximum					maximum
nClusters	.	.	.	.	40
algorithm	.	.	.	.	1
nDescTerms	.	.	.	.	15
spellCheck			N		
spellSensitivity			Med		
cellWeight			DEFAULT		
termWeight			DEFAULT		
minDocs	.	.	4	.	.
maxTerms			.		
resultTerms			ALL		
maxviewTerms	.	20000	20000	.	.
spellDict					
searchPhrase					
searchVar					
whereDoc					
lastfilternode					
lastparsenode					
synonymImport					
saveSynDS					
spellDS					
interStopDS					
interSynDS					
filters					
language		ENGLISH			
delimit		Std			
bCapitalize		Y			
bPartOfSpeech		Y			
NounGroups		Y			
multiDS		SASHELP.ENG_MULTI			
bPatterns		NONE			
stopList		SASHELP.ENGSTOP			
ignorePOS		'AUX' 'CONJ' 'DET' 'INTERJ' 'PART' 'PREP' 'PRON'			
ignoreAttrib		'NUM' 'PUNCT'			
bStems		Y			
synonymDS		SASHELP.ENGSYNMS			
filterLang					
TGConcepts					
TGCategories					
ignoreEntities					
startList					

parseVar		para			
topTermCnt	.	.	.	0	.
autoTopicCnt	.	.	.	25	.
autoTopic				N	
initTopics					
tm_topic_node					
tm_topic_dataset					
topics					
augTopics					

Appendix 3 - Stoplist Terms

'd, 'll, 'm, 're, 's, 've, a, aboard, about, above, according, accordingly, across, actually, after, afterwards, again, against, ago, ah, ain, all, almost, along, alongside, already, also, although, altogether, am, amid, amidst, among, amongst, an, and, another, any, anybody, anyhow, anyone, anyplace, anything, anyway, anyways, anywhere, apart, appreciate, appropriate, are, around, as, aside, ask, asking, associated, at, atop, available, away, b, be, became, because, become, becomes, becoming, been, before, behind, being, believe, below, beneath, beside, besides, between, beyond, both, but, by, c, call, called, came, can, can't, certain, certainly, change, changes, co, com, come, concerning, consequently, consider, considering, corresponding, could, course, currently, d, describe, described, despite, did, didn't, do, does, doesn't, doing, don't, done, during, e, each, edu, eg, either, else, enough, et, etc, etc., even, ever, every, everybody, everyone, everything, everywhere, example, except, f, following, for, former, formerly, forth, from, furthermore, g, generally, get, gets, getting, give, given, gives, go, goes, going, gone, got, gotten, greetings, h, had, hadn't, hardly, has, have, he, he'd, he's, hello, hence, her, here, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, i, i'd, i'm, ie, if, in, inasmuch, indeed, inside, insofar, instead, into, is, isn't, it, it's, its, itself, j, just, k, l, least, lest, let, let's, like, ll, m, ma'am, madam, made, make, many, may, maybe, me, meanwhile, merely, might, mine, minus, more, moreover, most, mr, mr., mrs, mrs., ms, ms., much, must, my, myself, n, n't, name, namely, nd, near, need, neither, never, nevertheless, new, next to, no, nobody, non, none, nonesuch, nonetheless, noone, nor, normally, not, nothing, notwithstand, now, o, obviously, of, off, often, oh, ok, okay, on, once, one, one's, only, onto, or, other, others, ought, our, ours, ourselves, out, out of, over, own, p, part, per, perhaps, please, plus, possible, presumably, probably, provide, provides, put, q, que, quite, qv, r, rather, rd, re, really, reasonably, regard, regarding, regardless, regards, regularly, relatively, respect, respectively, s, said, same, say, saying, says, secondly, see, seem, seemed, seeming, seems, self, selves, send, sensible, sent, seriously, shall, she, she'd, should, since, sir, so, some, somebody, somehow, someone, someplace, something, sometime, sometimes, someway, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, surely, t, take, taken, tell, tend, tends, th, than, thank, thanks, thanx, that, that's, thats, the, their, theirs, them, themselves, then, thence, there, there's, thereafter, thereby, therefore, therein, theres, thereupon, these, they, they're, think, this, thorough, thoroughly, those, though, through, throughout, thru, thus, thusly, thx, till, to, together, too, took, toward, towards, tried, tries, truly, try, trying, two, u, un, under, underneath, unfortunately, unless, unlike, until, unto, up, upon, us, use, used, useful, uses, using, usually, uucp, v, various, ve, very, via, viz, vs, w, was, wasn't, way, we, we'll, welcome, went, were, what, whatever, when, whence, where, whereafter, whereas, whereby, wherein, whereupon, wherever, wherewithall, whether, which, whichever, while, whilst, whither, who, whoever, whole, whom, whomever, whose, whosoever, why, will, willing, with, withal, within, without, won, won't, wonder, would, wouldn't, x, y, y'all, yet, you, you'll, you're, you've, your, yours, yourself, yourselves, z,

Appendix 4 – A Codebook for the Topics2 Dataset

(embedded as a pdf – should be restructured for the document)

Codebook for SAS Dataset: TOPICS2

Dataset

Date Created:	2014-11-07T13:43:19.3
Date Last Modified:	2014-11-07T13:43:19.4
Number of Observations:	25
Number of Variables:	8
Encoding:	wlatin1 Western (Windows)
Engine:	V9

_____extended attributes_____

Abstract:	This dataset was created from the October 20-22 minutes of the NSF1448107 sponsored group attending Dagstuhl event 14432. Topics were generated using default settings of SAS Text Miner on the concatenated raw minutes files, one record per paragraph, for the first three days of the meeting.
AccessRights:	Freely available, with attribution
Contributor:	Mary Vardigan(conceptualization, equal), Sam Hume(conceptualization, equal), Sanda Ionescu(conceptualization, equal), Jay Greenfield(conceptualization, equal), Jeremy Iverson(conceptualization, equal), John Kunze(conceptualization, equal), Barry Radler(conceptualization, equal), Stuart Weibel(conceptualization, equal), Michael C. Witt(conceptualization, equal)
Creator:	Larry Hoyle
Description:	This dataset is intended as an example for attaching source information to a dataset and a variable.
FundingInformation:	This dataset was created during Dagstuhl event 14432 by a group funded from NSF grant number1448107.
Language:	en-US
License:	Freely available, with attribution
Permanence:	Permanent: Unchanging Content
PublicationDate:	2014-11-17
Publisher:	University of Kansas
RelatedResourceAuthor:	Jay Greenfield, Larry Hoyle, Sam Hume, Sanda Ionescu, Jeremy Iverson, John Kunze, Barry Radler, Mary Vardigan, Stuart Weibel, Michael C. Witt
RelatedResourcePublicationDate:	2014-11-17
RelatedResourcePublisher:	University of Kansas
RelatedResourceRelationship:	isDerivedFrom
RelatedResourceTitle:	Minutes for Oct 20-22 2014 from NSF1448107 group at Dagstuhl event 14432
ResourceType:	dataset
SpatialCoverage:	Schloss Dagstuhl, Wadern, Germany
Study_AnalysisUnit:	paragraphs from raw minutes files
Study_CollectionMethodology:	Minutes were generated as Google Docs, one for each day at the Dagstuhl workshop. All participants could simultaneously edit the daily minutes file. Minutes for 2014-10-20, 2014-10-21, and 2014-10-22were copied from downloaded Microsoft Word files and concatenated into a single text file. This file was read into SAS and then used as input for SAS Text Miner with all default options chosen. The Topics results table was exported as this SAS dataset
Study_FundingInformation:	Participant travel and accomodations funded by NSF grant 1448107
Study_KindOfData:	SAS Text Miner Topics results table. Derived topics descriptions, dataset includes metadata in SAS extended attributes
Study_ProcessingDescription:	/* data read from concatenated minutes */ filename mins "C:\DDRIVE\projects\various\DDI\NSFDearColleague\MineMinutes\Oct20_22Minutes.txt"; libname minlib "C:\DDRIVE\projects\various\DDI\NSFDearColleague\MineMinutes"; data minlib.minutes; infile mins lrecl=520 pad; input para $520.; run; /* Dataset CITE.Topics

Appendix 5 – Extended Attributes for the Dataset

Alphabetic List of Data Set Extended Attributes

Extended Attribute	Character Value
Abstract	This dataset was created from the October 20-22 minutes of the NSF1448107 sponsored group attending Dagstuhl event 14432. Topics were generated using default settings of SAS Text Miner on the concatenated raw minutes files, one record per paragraph, for the first three days of the meeting.
AccessRights	Freely available, with attribution
Contributor	Mary Vardigan(conceptualization, equal), Sam Hume(conceptualization, equal), Sanda Ionescu(conceptualization, equal), Jay Greenfield(conceptualization, equal), Jeremy Iverson(conceptualization, equal), John Kunze(conceptualization, equal), Barry Radler(conceptualization, equal), Stuart Weibel(conceptualization, equal), Michael C. Witt(conceptualization, equal)
Creator	Larry Hoyle
Description	This dataset is intended as an example for attaching source information to a dataset and a variable.
FundingInformation	This dataset was created during Dagstuhl event 14432 by a group funded from NSF grant number1448107.
Language	en-US
License	Freely available, with attribution
ParentDatasets	Cite.Minutes, Cite.Topics
Permanence	Permanent: Unchanging Content
PublicationDate	11/17/2014
Publisher	University of Kansas
RelatedResourceAuthor	Jay Greenfield, Larry Hoyle, Sam Hume, Sanda Ionescu, Jeremy Iverson, John Kunze, Barry Radler, Mary Vardigan, Stuart Weibel, Michael C. Witt
RelatedResourcePublicationDate	11/17/2014
RelatedResourcePublisher	University of Kansas
RelatedResourceRelationship	isDerivedFrom
RelatedResourceTitle	Minutes for Oct 20-22 2014 from NSF1448107 group at Dagstuhl event 14432
ResourceType	dataset
SpatialCoverage	Schloss Dagstuhl, Wadern, Germany
Study_AnalysisUnit	paragraphs from raw minutes files
Study_CollectionMethodology	Minutes were generated as Google Docs, one for each day at the Dagstuhl workshop. All participants could simultaneously edit the daily minutes file. Minutes for 2014-10-20, 2014-10-21, and 2014-10-22were copied from downloaded Microsoft Word files and concatenated into a single text file. This file was read into SAS and then used as input for SAS Text Miner with all default options chosen. The Topics results table was exported as this SAS dataset
Study_FundingInformation	Participant travel and accomodations funded by NSF grant 1448107
Study_KindOfData	SAS Text Miner Topics results table. Derived topics descriptions, dataset includes metadata in SAS extended attributes

Study_ProcessingDescription	/* data read from concatenated minutes */ filename mins "C:\DDRIVE\projects\various\DDI\NSFDearColleague\MineMinutes\Oct20_22Minutes.txt"; libname minlib "C:\DDRIVE\projects\various\DDI\NSFDearColleague\MineMinutes"; data minlib.minutes; infile mins lrecl=520 pad; input para $520.; run; /* Dataset CITE.Topics generated from SAS Text Miner from minlib.minutes, CIte.Topics2 then generated by */ data CITE.TOPICS2; set CITE.TOPICS; Length TopicDescription $ 1000; TopicDescription = catx(" ","Topic ",_topicID," has ",_numDocs," documents and",_name," as Terms:"); run;
Study_Purpose	a sample dataset for enhanced data citation
TemporalCoverage	2014-10-20 to 2014-10-22
Title	Topics generated from minutes from NSF1448107 group at Dagstuhl event 14432
TopicalCoverage	Enhanced citation
Version	1
VersionDate	11/17/2014
VersionResponsibility	Larry Hoyle

Notes:

The only available UserID is the name of the dataset
Combining Description and Abstract gives a more complete description
No Subtitle or AlternateTitle

Appendix 6 – Extended Attributes for the Variable *TopicDescription*

Alphabetic List of Extended Attributes on Variables

Extended Attribute	Attribute Variable	Character Value
AccessRights	TopicDescription	Freely available, with attribution
AnalysisUnit	TopicDescription	paragraphs
Concept	TopicDescription	A label for a topic generated by SAS Text Miner combining the topic number, the number of documents relating to teh topic and the key descriptive terms for the document.
Contributor	TopicDescription	Mary Vardigan(writing – review & editing, lead)
Creator	TopicDescription	Larry Hoyle
Description	TopicDescription	A variable to be used with a Topics results dataset produced by SAS Enterprise Miner Test Miner. One string includes topic number, number of related Documents, and key terms.
GenerationInstruction	TopicDescription	TopicDescription = catx(" ","Topic ",_n_," has ",_numDocs," documents and",_name," as Terms:");
Language	TopicDescription	en-US
LevelOfMeasurement	TopicDescription	Nominal
Permanence	TopicDescription	Permanent: Unchanging Content
ProcessingDescription	TopicDescription	computed with the following SAS assignment statment: TopicDescription = catx(" ","Topic ",_topicID," has ",_numDocs," documents and",_name," as Terms:"); _topicID, _numDocs, and _name are standard variable names from an unsorted Topics dataset saved from Enterprise Miner.
PublicationDate	TopicDescription	11/17/2014
Publisher	TopicDescription	University of Kansas
ResourceType	TopicDescription	Variable
Role	TopicDescription	Potentially useful for topic labeling
Title	TopicDescription	Topic Descriptor Combining Sequence Number, Number of Related Documents, and Terms List From A SAS Text Miner Text Topics Node Result Table
VariableIdentifier	TopicDescription	TopicDescription
Version	TopicDescription	1
VersionDate	TopicDescription	2014_10_23
VersionResponsibility	TopicDescription	Larry Hoyle

Notes:

No unique identifier outside of SAS dataset until DDI finalized
License information appears as AccessRights
No Subtitle or AlternateTitle