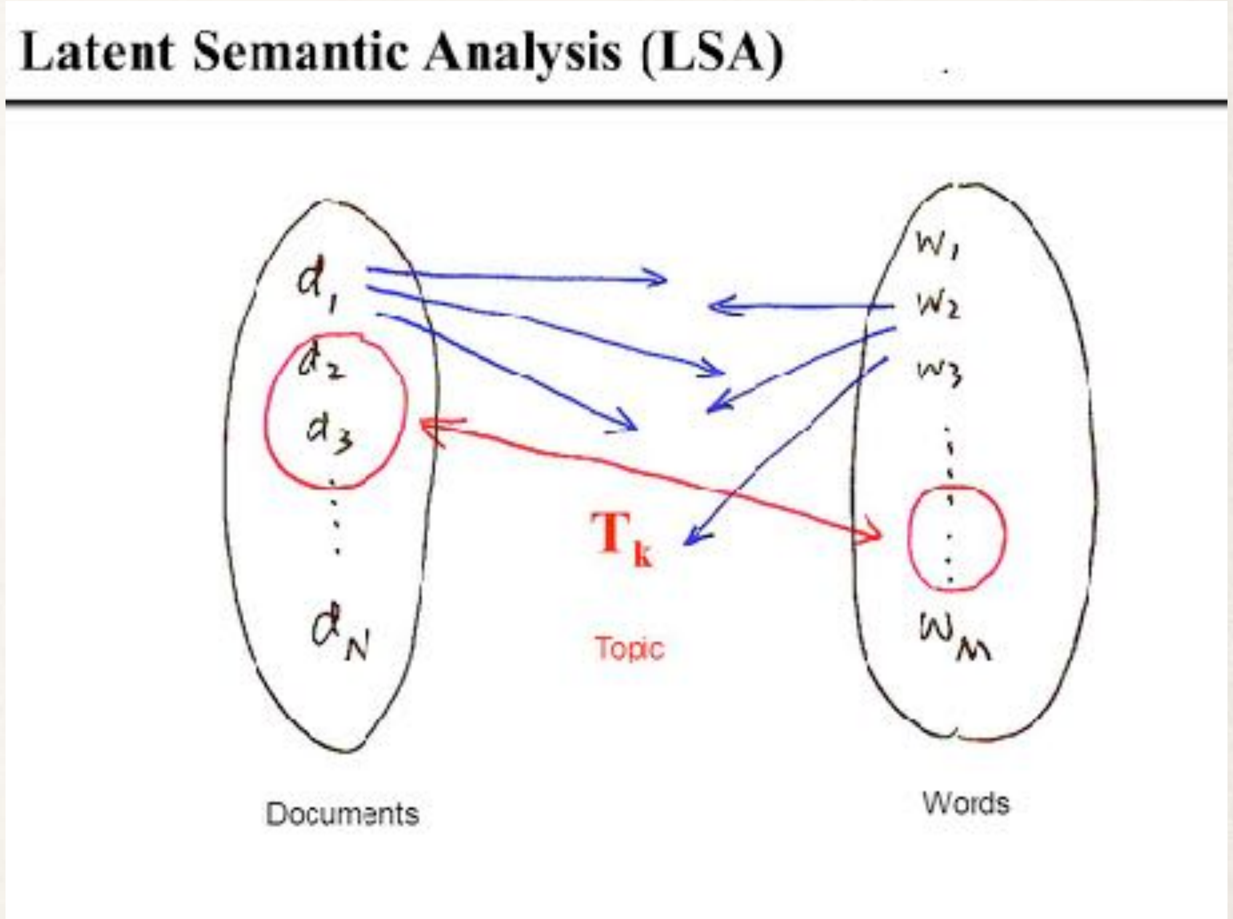
Topic Analysis of RDA Activities

A preliminary report

RDA TAB and IGWG Chairs Meeting, Gaithersburg, MD January 2017

Motivation

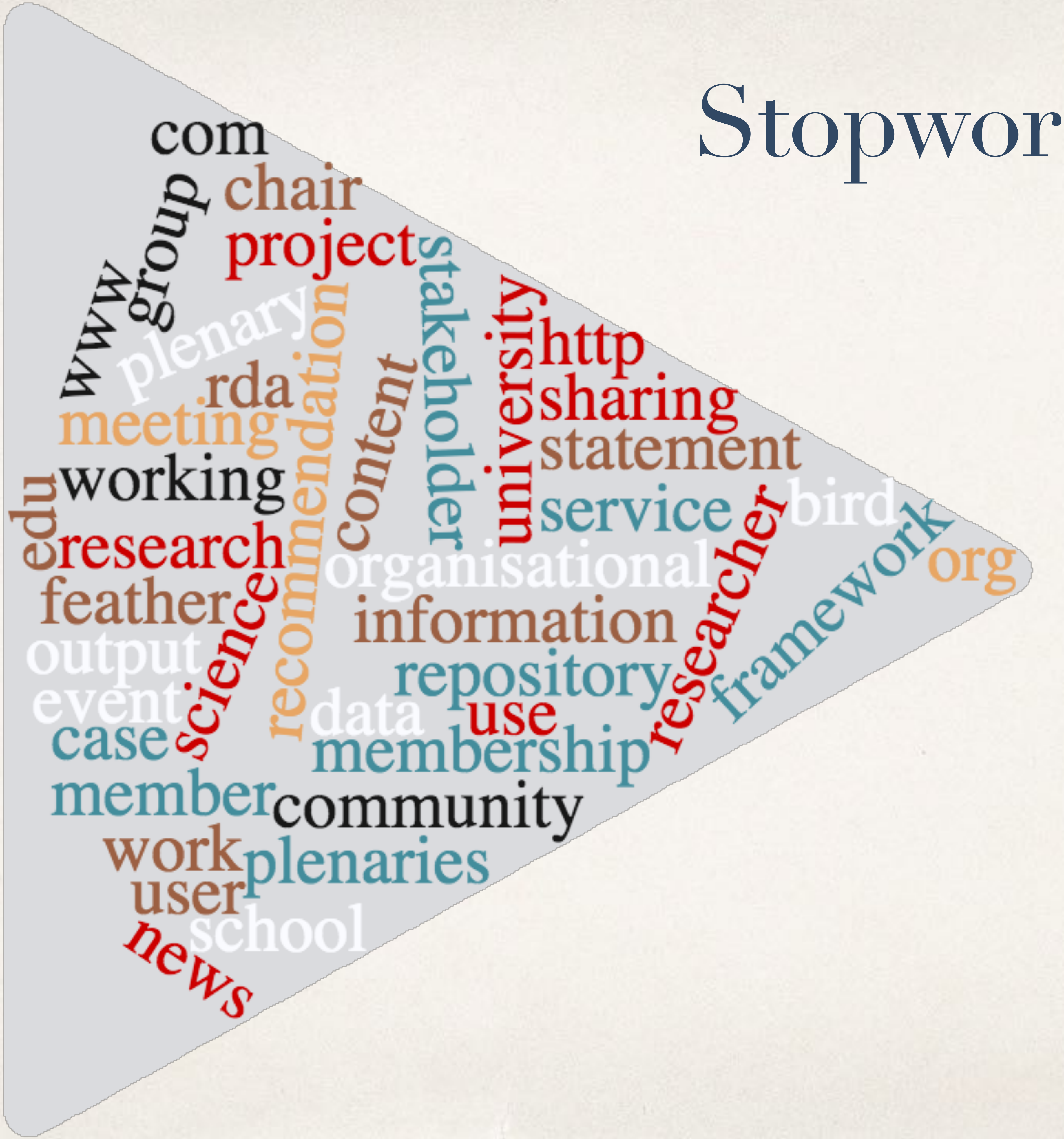
- Supplement the “Six Words” exercise
- Learn what RDA groups are doing by analyzing the documents they produce
- Create a corpus by crawling the RDA web pages
- Apply well-established techniques of Probabilistic Latent Semantic Analysis



Methodology

- Crawl rd-alliance.org using Apache Nutch, put docs in Solr (N=6,102)
- Select all documents containing "case-statement" (N=196)
- Remove stop words, do lemmatization
- Use *Gensim* toolkit to do Latent Dirichlet Allocation (LDA)

Stopwords


com
chair
group
www
project
stakeholder
university
plenary
recommendation
content
http
rda
sharing
meeting
statement
working
edu
service
bird
research
researcher
framework
org
organisational
feather
science
information
output
repository
event
data
use
case
membership
member
community
work
plenaries
user
school
news

Basic Concepts

- Topic Modeling identifies ***topics*** and their ***distributions*** across the ***documents*** in a ***corpus***.

- Generative probabilistic models use statistical methods to discover hidden (i.e., "latent") themes (topics) in documents.

- "Generative" means we assume the source documents were generated from a mixture of topics, with each topic being a distribution over words belonging to that topic.

- A word can belong to multiple topics.

- Neither the order of the documents in the corpus, nor the order of the words in each document are taken as significant.

LDA results w/ 15 topics, 8 words

Topic								
0	plan	platform	dmp	fabric	requirement	course	management	dmps
1	provenance	disciplinary	discipline	professor	reproducibility	illinois	collaboration	need
2	collection	privacy	pid	metadata	quality	digital	set	related
3	link	literature	infrastructure	publishing	article	publisher	scholix	hub
4	brokering	mediation	component	registry	resource	standard	provide	need
5	national	workflow	publishing	tool	publication	preservation	policy	activity
6	indigenous	sov	international	network	sovereignty	security	citation	trust
7	cost	database	field	recovery	farmer	centre	network	management
8	storage	qos	provider	datalc	vocabulary	multiple	document	adoption
9	legal	interoperability	law	domain	codata	international	national	ccm
10	type	creating	past	registry	standard	national	record	joining
11	agriculture	interoperability	agricultural	semantic	marine	semantics	infrastructure	vre
12	brokering	governance	model	infrastructure	middleware	interoperability	approach	support
13	metadata	standard	sample	practice	fishery	vocabulary	digital	humanity
14	rice	interoperability	ontology	policy	wheat	standard	organization	common

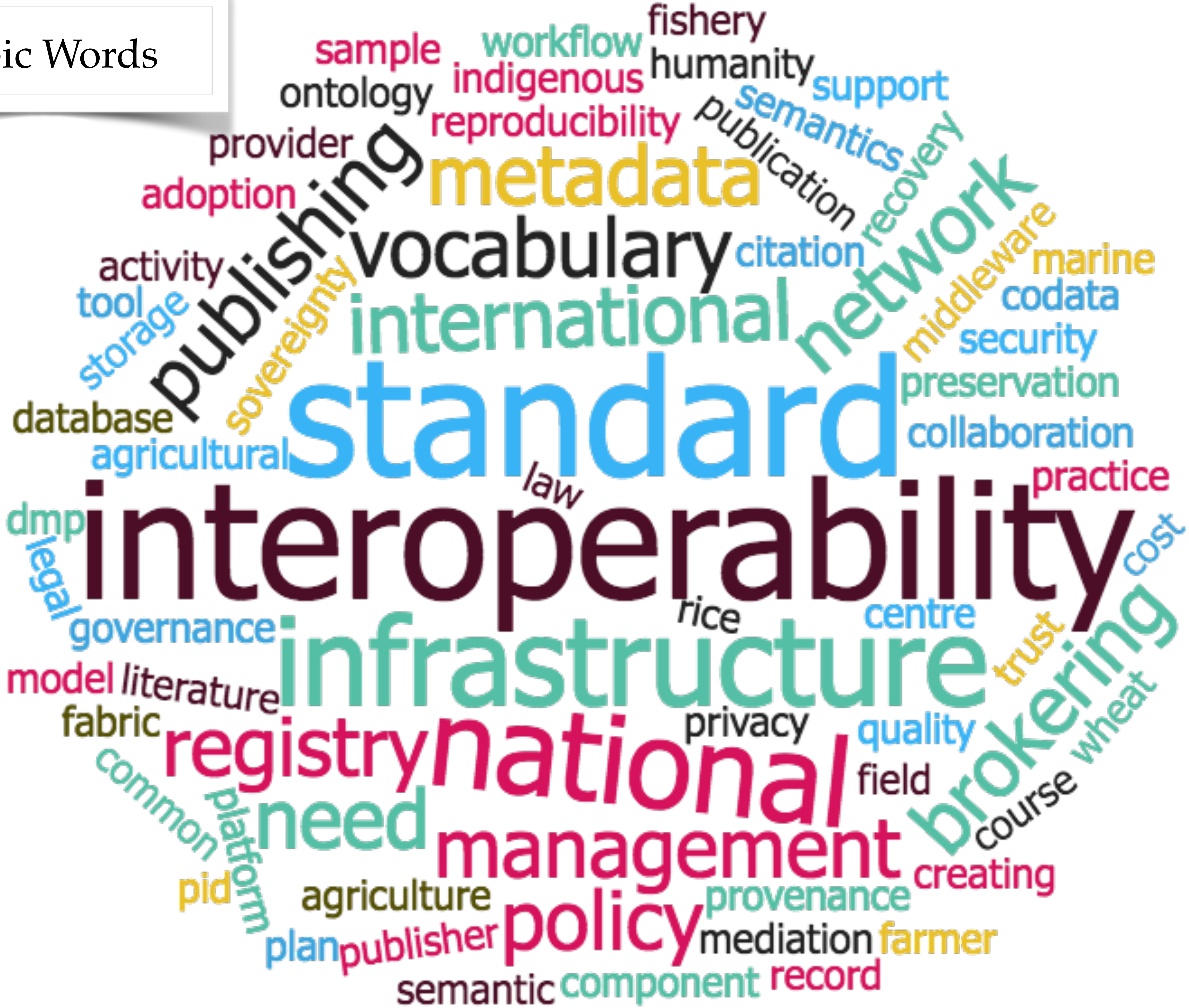
LDA results w/ 15 topics, 8 words

Topic								
0	plan	platform	dmp	fabric	requirement	course	management	dmps
1	provenance	disciplinary	discipline	professor	reproducibility	illinois	collaboration	need
2	collection	privacy	pid	metadata	quality	digital	set	related
3	link	literature	infrastructure	publishing	article	publisher	scholix	hub
4	brokering	mediation	component	registry	resource	standard	provide	need
5	national	workflow	publishing	tool	publication	preservation	policy	activity
6	indigenous	sov	international	network	sovereignty	security	citation	trust
7	cost	database	field	recovery	farmer	centre	network	management
8	storage	qos	provider	datalc	vocabulary	multiple	document	adoption
9	legal	interoperability	law	domain	codata	international	national	ccm
10	type	creating	past	registry	standard	national	record	joining
11	agriculture	interoperability	agricultural	semantic	marine	semantics	infrastructure	vre
12	brokering	governance	model	infrastructure	middleware	interoperability	approach	support
13	metadata	standard	sample	practice	fishery	vocabulary	digital	humanity
14	rice	interoperability	ontology	policy	wheat	standard	organization	common

LDA results w/ 15 topics, 8 words

Topic								
0	plan	platform	dmp	fabric	requirement	course	management	dmps
1	provenance	disciplinary	discipline	professor	reproducibility	illinois	collaboration	need
2	collection	privacy	pid	metadata	quality	digital	set	related
3	link	literature	infrastructure	publishing	article	publisher	scholix	hub
4	brokering	mediation	component	registry	resource	standard	provide	need
5	national	workflow	publishing	tool	publication	preservation	policy	activity
6	indigenous	sov	international	network	sovereignty	security	citation	trust
7	cost	database	field	recovery	farmer	centre	network	management
8	storage	qos	provider	datalc	vocabulary	multiple	document	adoption
9	legal	interoperability	law	domain	codata	international	national	ccm
10	type	creating	past	registry	standard	national	record	joining
11	agriculture	interoperability	agricultural	semantic	marine	semantics	infrastructure	vre
12	brokering	governance	model	infrastructure	middleware	interoperability	approach	support
13	metadata	standard	sample	practice	fishery	vocabulary	digital	humanity
14	rice	interoperability	ontology	policy	wheat	standard	organization	common

Topic Words


fishery
sample
workflow
humanity
indigenous
support
ontology
reproducibility
semantics
publication
provider
recovery
metadata
adoption
network
publishing
vocabulary
citation
marine
activity
middleware
codata
tool
international
storage
sovereignty
security
preservation
standard
database
collaboration
agricultural
law
practice
dmp
interoperability
cost
legal
rice
centre
governance
trust
infrastructure
brokering
model
literature
wheat
fabric
privacy
quality
registry
national
field
common
platform
need
course
management
creating
pid
agriculture
provenance
policy
plan
publisher
mediation
farmer
semantic
component
record

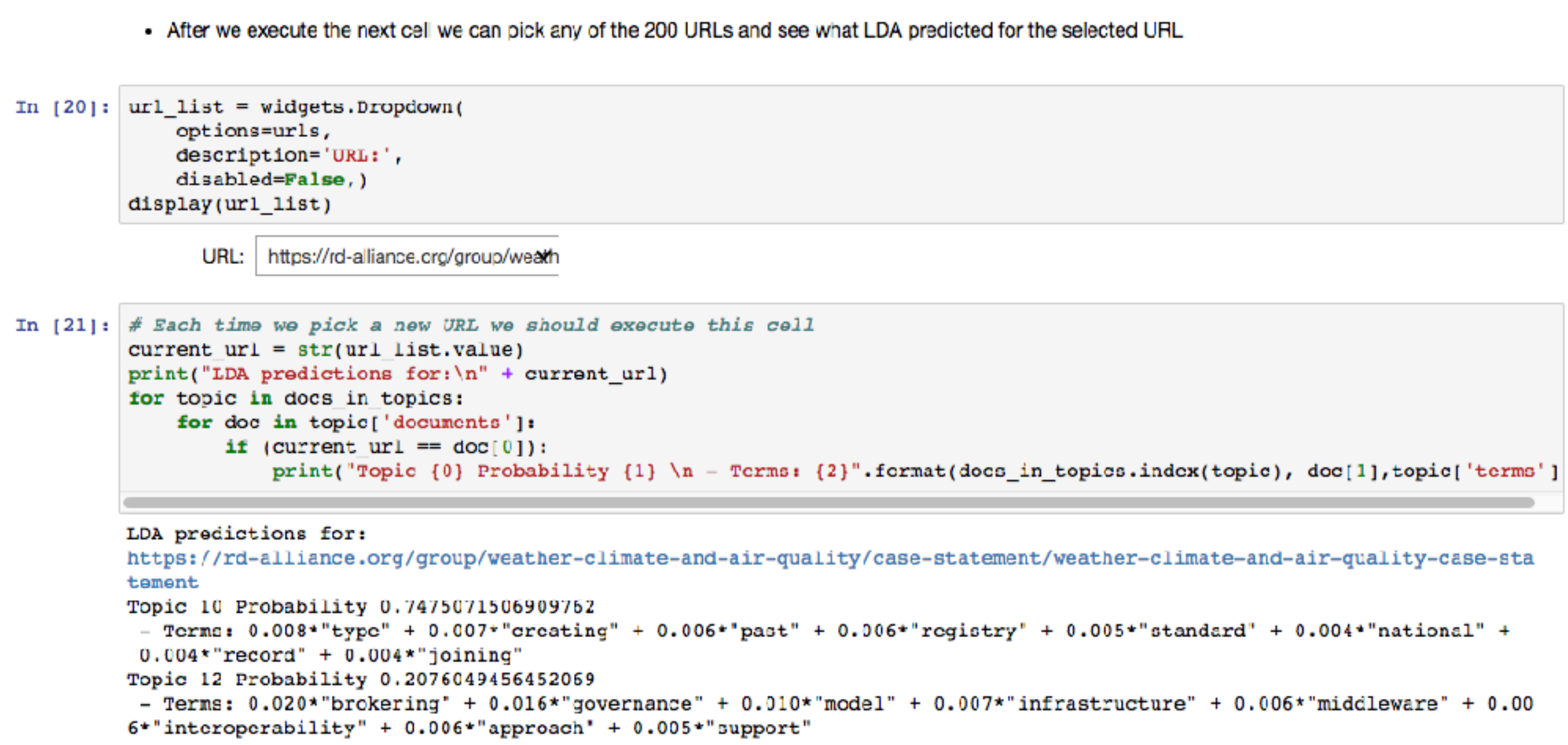
Sample Topics

- Case Statement from "Agrisemantics" WG has high probability (P=0.999) that it contains Topic 11:
  - *agriculture, interoperability, semantic, marine, semantics, infrastructure, vre*
  - Topic 11 also applies to "Virtual Research Environments" IG (P=0.997) and "Marine Data Management" WG (P=0.672)
- Statement for "On-Farm Data Sharing" WG contains Topic 7:
  - *cost, database, field, recovery, farmer, centre, network, management*
- Earlier LDA analysis had "Array Database" WG in Topic with:
  - *array, database, big, domain, document, report, plan*
  - but in this analysis has P=0.995 that contains Topic 7

Another Example

- Topic 7: ***indigenous, sov, international, network, sovereignty, security, citation, trust*** comes from these case statements:
    - International Indigenous Data Sovereignty IG (P = 0.999)
    - Data Security and Trust WG (P = 0.998)
    - Data for Development IG (P = 0.783)
    - Data Citation WG (P = 0.772)

Workflow in Jupyter Notebook



- After we execute the next cell we can pick any of the 200 URLs and see what LDA predicted for the selected URL

```
In [20]: url_list = widgets.Dropdown(
             options=urls,
             description='URL:',
             disabled=False,)
         display(url_list)
```

URL: https://rd-alliance.org/group/weath

```
In [21]: # Each time we pick a new URL we should execute this cell
         current_url = str(url_list.value)
         print("LDA predictions for:\n" + current_url)
         for topic in docs_in_topics:
             for doc in topic['documents']:
                 if (current_url == doc[0]):
                     print("Topic {0} Probability {1} \n - Terms: {2}".format(docs_in_topics.index(topic), doc[1],topic['terms']
```

```
LDA predictions for:
https://rd-alliance.org/group/weather-climate-and-air-quality/case-statement/weather-climate-and-air-quality-case-sta
tement
Topic 10 Probability 0.7475071506909762
 - Terms: 0.008*"type" + 0.007*"creating" + 0.006*"past" + 0.006*"registry" + 0.005*"standard" + 0.004*"national" +
 0.004*"record" + 0.004*"joining"
Topic 12 Probability 0.2076049456452069
 - Terms: 0.020*"brokering" + 0.016*"governance" + 0.010*"model" + 0.007*"infrastructure" + 0.006*"middleware" + 0.00
6*"interoperability" + 0.006*"approach' + 0.005*"support"
```

LDA results can be viewed at:

https://goo.gl/V6Yraj

What's Next?

- Refine list of stop words
- Remove duplicate documents (last version only?)
- Further tune algorithms
- Include other documents (e.g. outputs)
- Aggregate topic probabilities by group
- Assign dates to document, show Topic trends
- Other?

Thanks for your attention!