Analysis of Topic Trends in the Research Field of Digital Preservation

Soohyung Joo, School of Information Science, University of Kentucky

Introduction/Methods

This poster presents the preliminary results of topic trend analysis in digital preservation related research in the recent ten years (2008-2017). Text mining techniques were applied to explore topics that were studied in the area of digital preservation. For data collection, this study retrieved research publications concerning digital preservation from the Library, Information Science & Technology Abstracts database (LISTA). Research articles were retrieved from LISTA where the query “digital preservation” appeared in the search fields of title, subject terms, or author-supplied keywords. The search was limited to “scholarly journals,” “article,” and “English.” In total, 694 articles were retrieved from representative library and information science journals, such as “Preservation, Digital Technology & Culture,” “Library Hi Tech,” “Microform & Imaging Review,” and “International Journal on Digital Libraries.” Then, a text corpus consisting of article titles and abstracts was generated for text mining. The collected text underwent a series of pre-processing, including tokenization, elimination of stopwords, and stemming. The Potter stemmer algorithm was employed for stemming (tartarus.org/martin/PorterStemmer). Term frequency analysis and the Latent Dirichlet Allocation algorithm were employed to explore which topics were discussed in the research area of digital preservation.

Term Frequency Analysis

Table 1 presents top 100 terms in this research domain. In total, 52,650 tokens for unique 4,463 words were observed after removing stop-words.

Rank	Term	Freq.	Percent	Rank	Term	Freq.	Percent
1	digit	2256	4.28%	50	purpos	149	0.28%
2	preserv	1814	3.45%	52	base	144	0.27%
3	librari	830	1.58%	53	electron	142	0.27%
4	archiv	692	1.31%	54	state	140	0.27%
5	inform	555	1.05%	54	approach	140	0.27%
6	data	485	0.92%	56	current	134	0.25%
7	articl	449	0.85%	57	open	133	0.25%
8	collect	441	0.84%	57	relat	133	0.25%
9	research	437	0.83%	59	communiti	131	0.25%
10	paper	366	0.70%	60	describ	128	0.24%
11	institut	348	0.66%	61	creat	127	0.24%
12	project	329	0.62%	62	requir	125	0.24%
13	manag	326	0.62%	63	result	124	0.24%
14	access	326	0.62%	64	tool	121	0.23%
15	repositori	311	0.59%	65	focus	120	0.23%
16	discuss	298	0.57%	66	support	116	0.22%
17	develop	297	0.56%	67	import	114	0.22%
18	univers	277	0.53%	67	publish	114	0.22%
19	provid	273	0.52%	69	futur	113	0.21%
20	studi	261	0.50%	70	strategi	111	0.21%
21	record	247	0.47%	71	activ	109	0.21%
22	practic	244	0.46%	71	collabor	109	0.21%
23	content	230	0.44%	73	mani	107	0.20%
24	issu	222	0.42%	73	implement	107	0.20%
24	includ	222	0.42%	75	journal	106	0.20%
26	resourc	217	0.41%	75	standard	106	0.20%
27	nation	215	0.41%	77	program	105	0.20%
28	technolog	211	0.40%	78	initi	104	0.20%
29	challeng	210	0.40%	78	framework	104	0.20%
30	term	201	0.38%	78	social	104	0.20%
30	materi	201	0.38%	81	within	103	0.20%
32	present	195	0.37%	81	address	103	0.20%
33	system	194	0.37%	81	user	103	0.20%
34	work	193	0.37%	81	case	103	0.20%
35	metadata	191	0.36%	81	scienc	103	0.20%
36	web	190	0.36%	81	report	103	0.20%
37	servic	188	0.36%	87	avail	102	0.19%
37	need	188	0.36%	87	explor	102	0.19%
39	long	187	0.36%	89	plan	100	0.19%
40	model	184	0.35%	89	organ	100	0.19%
41	heritag	181	0.34%	89	scholar	100	0.19%
42	cultur	177	0.34%	92	identifi	98	0.19%
43	process	169	0.32%	92	govern	98	0.19%
44	curat	164	0.31%	94	method	96	0.18%
45	polici	158	0.30%	94	offer	96	0.18%
46	public	157	0.30%	96	survey	95	0.18%
47	author	156	0.30%	97	network	94	0.18%
48	document	154	0.29%	98	format	93	0.18%
49	find	153	0.29%	98	solut	93	0.18%
50	object	149	0.28%	100	review	91	0.17%

Table 1. Top 100 ranked terms

Table 2 shows the changes of top terms over ten years.

2008		2009		2010		2011		2012	
digit	4.48%	digit	4.52%	digit	4.84%	digit	4.86%	digit	5.09%
preserv	3.17%	preserv	3.34%	preserv	3.66%	preserv	3.36%	preserv	4.20%
librari	2.74%	librari	1.88%	archiv	2.25%	librari	2.20%	librari	1.24%
archiv	1.48%	inform	1.28%	librari	1.59%	research	1.29%	inform	1.13%
articl	1.37%	articl	1.21%	inform	1.20%	archiv	1.26%	archiv	0.99%
inform	1.11%	archiv	1.08%	paper	1.02%	repositori	1.21%	research	0.84%
collect	1.05%	project	0.97%	project	0.79%	univers	1.05%	articl	0.82%
univers	1.03%	repositori	0.94%	articl	0.74%	collect	0.94%	repositori	0.80%
project	0.84%	collect	0.94%	manag	0.74%	access	0.89%	paper	0.77%
access	0.84%	data	0.89%	record	0.72%	manag	0.86%	develop	0.73%
institut	0.75%	develop	0.74%	provid	0.69%	discuss	0.86%	data	0.69%
manag	0.73%	metadata	0.73%	access	0.67%	inform	0.78%	collect	0.69%
data	0.71%	discuss	0.72%	discuss	0.64%	institut	0.78%	manag	0.68%
repositori	0.69%	nation	0.71%	research	0.61%	paper	0.73%	access	0.64%
record	0.69%	institut	0.69%	data	0.56%	articl	0.70%	term	0.60%
electron	0.67%	access	0.65%	communiti	0.54%	data	0.70%	institut	0.57%
provid	0.67%	research	0.65%	resourc	0.54%	project	0.59%	process	0.57%
discuss	0.62%	web	0.59%	develop	0.54%	system	0.54%	provid	0.57%
materi	0.60%	materi	0.56%	issu	0.54%	resourc	0.54%	servic	0.57%
develop	0.60%	state	0.51%	cost	0.54%	issu	0.54%	polici	0.53%
nation	0.56%	univers	0.48%	includ	0.54%	scholar	0.54%	heritag	0.51%
servic	0.54%	manag	0.47%	challeng	0.51%	curat	0.51%	resourc	0.49%
web	0.52%	includ	0.47%	purpos	0.51%	technolog	0.48%	system	0.49%
paper	0.51%	cultur	0.47%	tool	0.51%	content	0.48%	project	0.49%
issu	0.51%	approach	0.46%	collect	0.49%	provid	0.48%	challeng	0.47%
content	0.49%	issu	0.46%	technolog	0.46%	challeng	0.46%	issu	0.47%
resourc	0.47%	need	0.44%	system	0.46%	cultur	0.46%	content	0.47%
studi	0.47%	provid	0.44%	object	0.44%	heritag	0.46%	discuss	0.46%
scholar	0.45%	heritag	0.44%	find	0.44%	museum	0.43%	long	0.46%
book	0.45%	program	0.43%	author	0.44%	publish	0.43%	model	0.44%

2013		2014		2015		2016		2017	
digit	4.42%	preserv	4.04%	digit	3.51%	digit	3.25%	digit	4.61%
preserv	3.17%	digit	3.71%	preserv	2.96%	preserv	2.72%	preserv	4.04%
inform	2.17%	research	1.37%	data	1.68%	data	1.45%	librari	1.22%
librari	1.33%	archiv	1.24%	archiv	1.60%	archiv	1.34%	archiv	1.14%
collect	1.28%	librari	1.16%	librari	1.21%	librari	1.28%	institut	1.07%
archiv	0.87%	institut	0.92%	research	0.94%	paper	1.03%	inform	1.05%
data	0.87%	paper	0.88%	articl	0.83%	collect	0.84%	paper	0.88%
studi	0.84%	collect	0.75%	collect	0.69%	research	0.82%	studi	0.78%
research	0.79%	inform	0.69%	inform	0.65%	curat	0.82%	collect	0.76%
project	0.79%	access	0.63%	practic	0.62%	inform	0.75%	research	0.74%
articl	0.77%	content	0.61%	web	0.56%	institut	0.73%	articl	0.69%
manag	0.69%	univers	0.61%	manag	0.55%	content	0.71%	access	0.69%
work	0.64%	articl	0.59%	discuss	0.54%	studi	0.63%	provid	0.65%
challeng	0.61%	practic	0.59%	paper	0.52%	technolog	0.63%	manag	0.61%
nation	0.61%	data	0.59%	provid	0.51%	develop	0.57%	data	0.61%
access	0.54%	repositori	0.59%	develop	0.50%	practic	0.57%	univers	0.59%
public	0.54%	manag	0.57%	servic	0.50%	articl	0.54%	polici	0.57%
technolog	0.54%	studi	0.55%	resourc	0.50%	record	0.54%	challeng	0.57%
develop	0.51%	web	0.51%	metadata	0.47%	manag	0.50%	project	0.55%
paper	0.51%	develop	0.47%	access	0.45%	resourc	0.50%	news	0.53%
practic	0.51%	heritag	0.47%	polici	0.44%	dspace	0.46%	discuss	0.53%
record	0.51%	includ	0.43%	present	0.44%	discuss	0.44%	work	0.51%
institut	0.49%	model	0.43%	project	0.43%	includ	0.44%	model	0.50%
polici	0.43%	nation	0.43%	need	0.41%	provid	0.42%	practic	0.48%
discuss	0.43%	educ	0.43%	includ	0.41%	univers	0.42%	develop	0.46%
comput	0.43%	discuss	0.43%	repositori	0.40%	base	0.42%	need	0.44%
framework	0.43%	present	0.43%	record	0.40%	project	0.42%	includ	0.44%
term	0.41%	record	0.43%	process	0.39%	model	0.42%	object	0.44%
provid	0.41%	work	0.41%	social	0.39%	open	0.40%	record	0.44%
materi	0.41%	challeng	0.41%	studi	0.39%	purpos	0.36%	purpos	0.42%

Table 2. Changes of top ranked terms from 2008 to 2017

LDA Topic Models

The LDA topic modeling uncovered 20 topics underlying the collected corpus. The findings revealed prevailing topics, ranging from archiving record management, cultural heritage preservation, digital preservation strategies, metadata development, web resource archiving, digital curation practice, institutional repositories, and to others. Also, Table 4 presents topics extracted from the recent four years (2014-2017) respectively.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
data	institut	inform	metadata	materi	resourc	digit	digit	paper	articl
research	repositori	technolog	develop	preserv	web	author	curat	purpos	includ
social	system	public	tool	educ	archiv	format	practic	find	present
scienc	plan	challeng	requir	legal	user	standard	work	issu	discuss
futur	provid	govern	sourc	media	increas	file	museum	provid	issu
manag	integr	need	describ	news	websit	document	art	practic	report
understand	implement	document	creat	current	page	pdfa	profession	digitis	topic
explor	applic	knowledg	process	law	number	born	asset	address	various
trust	mani	preserv	open	field	time	inform	life	originalityvalu	offer
develop	support	chang	need	high	evalu	guidelin	artist	literatur	introduct
Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
access	object	librari	preserv	preserv	collect	digit	studi	archiv	term
content	approach	univers	communiti	digit	project	heritag	polici	record	long
publish	base	book	collabor	strategi	articl	cultur	research	manag	servic
journal	model	academ	network	ensur	digit	nation	manag	archivist	solut
scholar	process	librarian	state	organis	focus	preserv	result	person	model
avail	framework	role	infrastructur	task	histori	problem	survey	discuss	cost
electron	specif	servic	support	start	special	manuscript	analysi	electron	Provid
discuss	propos	centuri	initi	action	imag	develop	method	author	Storag
articl	concept	googl	organ	train	select	countri	identifi	risk	sustain
copi	result	protect	build	issu	histor	intern	exist	apprais	cloud

Table 3. Research topics in the area of digital preservation: LDA topic model (k=20)

2017

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
studi	archiv	digit	paper	digit	artist	record	librari	technolog	preserv
practic	document	failur	access	articl	media	literatur	review	chang	institut
museum	discuss	mani	academ	object	book	person	law	content	challeng
curat	issu	addit	univers	organis	art	manag	platform	key	build
purpos	initi	import	staff	approach	strategi	phrs	includ	annot	knowledg
case	web	nation	histori	preserv	perspect	includ	call	infrastructur	work
framework	websit	provid	implic	offer	structur	exist	inform	explor	provid
T11	**T12**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**
news	term	inform	polici	digit	cultur	collect	method	model	materi
project	long	data	preserv	preserv	communiti	librari	find	assess	process
present	storag	research	univers	creat	solut	limit	tradit	exist	associ
fake	sever	manag	provid	find	collabor	develop	base	repositori	object
journal	resourc	need	institut	grow	open	technic	manuscript	identifi	present
meet	content	system	studi	report	develop	base	born	area	intern
result	growth	avail	digit	effect	heritag	focus	examin	self	tool

2016

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
articl	research	archiv	institut	model	offer	librari	process	preserv	data
histori	consid	born	manag	educ	metadata	academ	document	communiti	scienc
technolog	held	review	studi	content	evid	librarian	practic	digit	social
method	view	base	heritag	paper	group	servic	framework	cloud	spatial
materi	effect	cours	resourc	avail	confer	challeng	electron	comput	infrastructur
goal	virtual	apprais	cultur	activ	american	chang	propos	place	explor
studio	identifi	examin	public	number	interest	develop	technic	base	survey
T11	**T12**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**
inform	digit	archiv	newspap	content	discuss	preserv	collect	digit	dspace
technolog	pdfa	resourc	number	includ	journal	digit	requir	curat	librari
develop	format	web	collabor	preserv	within	paper	need	research	websit
issu	practic	javascript	support	articl	initi	open	nation	histor	adopt
creat	materi	respons	record	project	research	purpos	program	plan	onlin
term	author	load	texa	media	project	strategi	print	practic	indian
long	literatur	embed	state	current	develop	govern	today	intern	univers

2015

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
servic	digit	need	articl	resourc	research	inform	web	digit	digit
long	cloud	access	metadata	archiv	relat	process	page	librari	project
term	pdfa	irish	creat	damag	archaeolog	model	content	publish	heritag
develop	storag	explor	record	impact	studi	manag	approach	paper	cultur
institut	show	part	journal	time	practic	specif	archiv	issu	nation
futur	format	signific	challeng	embed	strategi	requir	chang	communiti	data
offer	author	establish	open	import	factor	implement	number	address	problem
T11	**T12**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**
digit	collect	preserv	document	materi	preserv	polici	data	discuss	archiv
preserv	librari	repositori	mission	onlin	content	archiv	manag	present	social
work	case	support	user	digit	literatur	target	geospati	includ	media
art	univers	integr	center	public	build	base	provid	report	studi
curat	result	system	govern	examin	infrastructur	select	share	topic	record
technolog	state	control	high	code	describ	allow	relev	editor	preserv
practic	includ	method	plan	various	interest	avail	organ	introduct	scienc

2014

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
model	preserv	research	provid	institut	librari	manag	work	preserv	preserv
project	paper	purr	need	repositori	technic	digit	object	inform	educ
relat	nation	repositori	environ	metadata	refer	heritag	document	communiti	field
award	effort	univers	integr	support	challeng	africa	techniqu	plan	address
deliv	practic	collabor	system	organ	role	find	includ	guidelin	program
understand	challeng	work	articl	resourc	cost	analysi	dcc	process	plenari
articl	design...	open	plan	explor	discuss	eastern	promot	discuss	high
T11	**T12**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**
research	activ	access	content	academ	digit	archiv	web	data	digit
digit	common	articl	format	student	open	record	method	collect	studi
develop	educ	includ	network	state	among	electron	result	univers	term
teach	knowledg	materi	mani	these	engin	import	increas	bangladesh	long
identifi	effort	legal	initi	manag	region	china	page	review	purpos
effect	learn	deposit	lockss	involv	made	studi	base	general	practic
toward	comment	public	discuss	solut	pradesh	fund	user	valu	special

Table 4. Research topics in recent years (2014-2017)