Attachment #4: IIA Rules

The following table represents the final rules used by IIA. The “generalizable” column provides an indication of whether these rules can be generalized to other agencies. The “Score” column provides the weight assigned to each rule, and whether they were positive or negative rules (positive rules are indicators of in scope documents and negative rules are indicators of documents that are not in scope. The “Attribute” column denotes what attributes were examined by the harvester for each rule, and “Values” are the actual words or phrases that were examined.

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
1.2		EPA-hosted Federal Register Notices			
1.2.1	S		-3	object-title	Federal Register, "For Immediate release
1.2.2	S		-3	keyword	Federal Register,"For Immediate release"
1.2.3	S		-3	document-summary	Federal Register,"For Immediate release"
1.2.4	S		-9	theurl	fedrgstr
1.2.5	S		-3	epa_breadcrumbs	Federal Register
1.2.6	S		-3	links_and_labels	Federal Register
1.2.7	S		-3	headings	Federal Register Notice
1.2.6	S		-3	highlighted	Federal Register
1.3		EPA news releases			
1.3.1	S		3	document-text	@epa.gov
1.3.2	S		2	document-text	for immediate release
1.3.2	S		2	document-text	for immediate release
1.4		EPA approved content			
1.4.1	S		1	document-text	epa approved,"epa has approved"
1.4.2	S		3	links_and_labels	epa approved,"epa has approved"
1.4.3	S		3	object-title	epa approved,"epa-approved","epa has approved"
1.5		Letters			
1.5.1	Y		-3	document-text	dear,"sincerely","thank you"
1.6		Procurement Office			
1.6.1	Y		-3	document-text	Procurement Office

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
2		Official reports			
2.1	Y	PDF reports			
2.1.1		PDF	+	Object-type	application/pdf
2.1.2					
2.1.2.1	S		2	document-text	fact sheet, "copies of this report available from", "copies of this fact sheet available from","List of Tables", "List of Images", "Table of Contents", "Environmental Impact Statement", EIS, "Environmental Impact report", EIR, "Request a hard copy", "Resource guide", "Technical Report", "Working paper", "Review Literature", "intentionally left blank"
2.1.2.2	Y		1	document-text	report,contents,introduction,references,revised
2.1.3	S		3	document-summary	Final report, "Fact Sheet", "Environmental Impact Statement", Proceedings
2.1.4	S		3	object-title	epa order,"Final report", "Fact Sheet", "Environmental Impact Statement", Proceedings
2.1.6	Y		-5	object-title	Draft
2.1.7	Y		-3	referrer_url	draft
2.1.8	S		3	Webi_description	Final report, "Fact Sheet", "Environmental Impact Statement", Proceedings
2.1.9	S		3	Webi_title	Final report, "Fact Sheet", "Environmental Impact Statement", Proceedings
2.1.10	Y		-3	Webi_title	Draft
2.1.11	Y		1	highlighted	fact sheet
2.1.12	S		1	links_and_labels	epa order
2.1.13		Child pages fact sheet			
2.1.13.1	Y		1	document-text	fact sheet
2.1.13.2	Y		1	object-title	fact sheet
2.1.13.3	Y		1	object-title	fact sheet

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
2.2		HTML reports			
2.2.1	Y	HTML	+	Object-type	text/html
2.2.2					
2.2.2.1	S		2	document-text	fact sheet, "copies of this report available from", "copies of this fact sheet available from", "List of Tables", "List of Images", "Table of Contents", "Environmental Impact Statement", EIS, "Environmental Impact report", EIR, "Request a hard copy", "Resource guide", "Working paper", "Review Literature", "Study Purpose", "Funding organization","intentionally left blank"
2.2.2.2	Y		1	document-text	report,contents,Introduction,references,revised
2.2.3	S		3	document-summary	Report, "fact sheet", "copies of this report available from", "copies of this fact sheet available from", Introduction, Content, References, "List of Tables", "List of Images", Attachments, "Table of Contents", "Environmental Impact Statement", EIS, "Environmental Impact report", EIR, "Proceedings of", "Request a hard copy", "Resource guide", "Working paper", Revised, "Review Literature", "Study Purpose", "Funding organization", "Funding provided by",
2.2.4	S		3	Keyword	fact sheet, "copies of this report available from", "copies of this fact sheet available from", Introduction, Content, References, "List of Tables", "List of Images", Attachments, "Table of Contents", "Environmental Impact Statement", "Environmental Impact report", EIR, "Proceedings of", "Request a hard copy", Resource guide", "Working paper", Revised, "Review Literature", "Study Purpose", "Funding organization", "Funding provided by",

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
2.2.11	S		3	Webi_description	Report, "fact sheet", "copies of this report available from", "copies of this fact sheet available from", Introduction, Content, References, "List of Tables", "List of Images", Attachments, "Table of Contents", "Environmental Impact Statement", EIS, "Environmental Impact report", "Proceedings of", "Request a hard copy", Resource guide", "Technical Report", "Working paper", Revised, "Review Literature", "Study Purpose", "Funding organization", "Funding provided by",
2.2.12	S		-3	Webi_title	Draft
2.2.13	Y		1	highlighted	fact sheet
2.1.14	N		1	links_and_labels	epa order
2.1.15			3	object-title	epa order,"Final report", "Fact Sheet", "Environmental Impact Statement", Proceedings
2.2.16	S	Child pages fact sheet			
2.2.16.1			1	document-text	fact sheet
2.2.16.2	Y		1	object-title	fact sheet
2.2.16.3	Y		1	object-title	fact sheet

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
3		EPA Posters			
		EPA Posters			
3.1		Descriptive rules			
3.1.1	Y		3	Object-type	image,media-video
3.1.2	Y		3	referrer_url	poster
3.1.3	Y		3	img_alt	Poster
3.1.4	Y		2	Highlighted	poster
4		EPA Program Descriptions			
		EPA Program Descriptions			
4.1		Keyword			
4.1.1	Y		2	document-text	official business,"program report"
4.1.2	Y		3	object-title	official business,"program report","program update"
4.1.3	Y		3	links_and_labels	official business,"program report","program update"
4.1.4	Y		2	Metadata	Geographic Area, "Project Officer"
4.1.5			3	Webi_title	official business,"program report","program update"
5		EPA Publications			
5.1		EPA Publications listed at NEPIS			
5.1.1	N		3	referrer_url	nepis.epa.gov/pubtitle
5.2		Publications by EPA researchers			
5.2.1	S		2	document-text	Source Document,"Agency Work Group Review", "Verification Date", "EPA Contacts", "Supporting Studies", "Quantitative Estimate", "EPA Documentation"

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
6		Agency Orgcharts			
6.1		Agency Orgcharts Features			
6.1.1	Y		3	Object-type	text/html, application-acrobat-pdf
6.1.2	Y		3	object-title	Organization* Chart, Organization
6.1.3	Y		3	document-summary	Organization* Chart, Organization, "Office of",
6.1.4	Y		2	Keyword	Organization* Chart, Organization. "Office of",
6.1.5	Y		2	Headings	Organization* Chart, Organization. "Office of",
6.1.6	Y		3	links_and_labels	Organization Chart, Organization, "Office of",
6.1.7	Y		3	img_alt	Organization Chart, Organization, "Office of",
6.1.7	Y		3	img_alt	Organization Chart, Organization, "Office of",
6.1.8	Y		3	Webi_keywords	Organization Chart, Organization
6.1.9	Y		1	Webi_description	Organization* Chart, Organization
6.2		Known related sources			
6.1.7			1	img_alt	Organization Chart, Organization, "Office of",
7		Agency Press Releases			
7.1		Indicators			
7.1.1	Y		1	document-text	Press Release
7.1.2	Y		3	object-title	Press Release
7.1.3	Y		3	labels	Press Release
7.1.4	Y		3	Keyword	Press Release
7.1.5	Y		3	links_and_labels	News Releases feed, "in the news"
7.1.6	Y		3	Webi_keywords	Press Release
7.1.7	Y		3	Webi_title	Press Release

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
7.2		Known agency subsystems containing Press Releases and Related Publications			
7.2.1	N		3	theurl	gov/newsroom/newsreleases ,opa/admpress
7.3		Public Service Announcements			
7.3.1			3	document-text	Public Service Announcement, PSA
7.3.2	Y		3	object-title	Public Service Announcement, PSA
7.3.3	Y		3	links_and_labels	Public Service Announcement, PSA
7.3.4	Y		2	Object-type	application-audio-mp3, text/html, application-adobe-pdf
7.3.5	N		3	referrer_url	gov/emergenc/katrina/outreach
7.3.6	Y		3	Webi_title	Public Service Announcement, PSA
8		Agency Advisories and Bulletins			
8.1		Advisories Indicators			
8.1.1	Y		3	object-title	Advisory on, "Advisory by"
8.1.2	Y		3	links_and_labels	Advisory on, "Advisory by"
8.1.3	Y		2	Headings	Advisory on, "Advisory by"
8.1.4	Y		1	document-text	Availability, Committee, Chair, "Table of contents", Abstract, "charge to subcommittee", "response by subcommittee"
8.1.5	Y		3	Webi_title	Advisory on, "Advisory by"

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
9		Funding Opportunities Announcements			
9.1		Funding Op. Indicators			
9.1.1	S		2	document-text	Solicitation, "Opening Date", "Closing Date", Eligibility, Submissions, "Application Form*", "Synopsis of Program", "funding opportunity", "award information", "under a grant", "federal grant","cooperative agreement", "form 424"
9.1.2	S		2	Headings	Solicitation, "Opening Date", "Closing Date", Eligibility, "Technical Contact", Submissions, "Application Form*", "Synopsis of Program", "funding opportunity", "award information", "under a grant", "federal grant","cooperative agreement", "standard form 424"
10		Environmental Indicators Documents			
11		Datasets and models			
11.1		Dataset indicators			
11.1.1	Y		1	Object-type	application-executable, media-archive, text/text
11.1.2	S		3	EIMS_Information_Type	Model,Dataset
12		EPA official responses to public comments			
12.1		ADI Control numbers			
12.1.1	S		3	ADI_Control_Number	*

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
13		Official Directives			
13.1	Y		3	object-title	guidance,"reporting guide"
14 pivot=		Parent-child rules			
14.1		Parent pages			
14.1.1	Y		3	links_and_labels	Chapter (\d+),"Section (\d+)","Appendix","Introduction","Cover","Findings"
14.2		Leaf pages			
14.2.1	Y		-3	headings	Chapter (\d+),"Section (\d+)"
14.2.2	Y		-3	object-title	Chapter (\d+),"Section (\d+)"
14.2.3	Y		-3	highlighted	Chapter (\d+),"Section (\d+)"
14.2.4	Y		-3	headings	Section (\d+),"Section (\d+)"
14.2.5	Y		-3	object-title	Section (\d+),"Section (\d+)"
14.2.6	Y		-3	highlighted	Section (\d+),"Section (\d+)"
14.2.7	Y		-3	headings	Introduction,"Cover","Findings"
14.2.8	Y		-3	object-title	Introduction,"Cover","Findings"
14.2.9	Y		-3	highlighted	Introduction,"Cover","Findings"
14.2.10	Y		-3	headings	Introduction,"Cover","Findings"
14.2.11	Y		-3	object-title	Introduction,"Cover","Findings"
14.2.12	Y		-3	highlighted	Introduction,"Cover","Findings"
14.3		Links or text			
14.3.1	Y	Mostly Links	3	@mostlylinks	
14.3.2	Y	Mostly Text	-3	@mostlytext	
15		Supplemental rules			
15.1	Y	Mostly Links	-6	@links	
15.2	N		3	theurl	airtrends/aqtrnd96/general
15.3	N		20	theurl	ttn/naaqs/ozone/areas
15.4	N		20	theurl	airmarket/emissions/raw/data
15.5	Y		-6	headings	permit
15.6	Y		-6	theurl	permit

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
15.6	Y		-6	object-title	permit
15.6	Y		-6	object-title	permit
21		Special handling			
21.1		Copyrighted material - special handling			
21.1.1	Y		3	document-text	copyright, "All rights reserved","Authorized use only","Not for distribution"
21.1.2	Y		3	copyright	*
21.2		Internal agency - special handling			
21.2.1	Y		3	headings	For official use,"For internal use","For administrative use","For operational use"
21.2.2	Y		2	document-text	For official use,"For internal use","For administrative use","For operational use"
21.3		Classified or restricted - special handling			
21.3.1	Y		3	headings	Classified,restricted
21.4		Documents with disclaimers			
21.4.1	Y		3	links_and_labels	Disclaimer
21.4.2	Y		2	Highlighted	Disclaimer
21.4.3	Y		1	document-text	Disclaimer
22		Works in progress			
22.1	Y		3	document-text	this is a test,"do not publish","limited distribution","not for distribution"
22.2	Y		3	theurl	text

RuleID	Generalizable y=yes,n=no, s=substitution	Description	Score	Attribute	Values
30		System-generated rules			
30.1		Rules applied to problem docment			
30.1.2			1	epa_breadcrumbs	water
30.1.3			1	epa_breadcrumbs	great lakes
30.1.4			1	epa_breadcrumbs	publications
30.1.5			1	epa_breadcrumbs	document
30.1.6			1	epa_breadcrumbs	system
30.1.7			1	epa_breadcrumbs	pollution
30.1.8			1	epa_breadcrumbs	environmental publications
30.1.9			1	epa_breadcrumbs	nepis
30.1.10			1	epa_breadcrumbs	toxics strategy
30.1.11			1	epa_breadcrumbs	science
30.1.12			1	theurl	ord/webpubs
30.1.13			1	theurl	projsum
30.1.14			1	theurl	safewater
30.1.15			1	tssms	download
30.1.16			1	tssms	safewater
30.1.17			1	epa_breadcrumbs	information
30.1.18			1	tssms	Clariton
30.1.19			1	object-title	epa (\d+)
30.1.20			1	subject	innovative hazardous waste
30.1.21			1	theurl	criteria
30.1.22			1	author	office of water
30.1.23			1	theurl	swertio1
30.1.24			1	object-title	drinking
30.1.26			-1	labels	(\d+)
30.1.27			-1	object-title	region
30.1.30			-1	referrer_url	yosemite.epa.gov/r10

RuleID	Description	Score	Attribute	Values
30.1.31		-1	referrer_url	air
30.1.32		-1	links_and_labels	(\d+)
30.1.35		-1	links_and_labels	yosemite.epa.gov/r10
30.1.37		-1	links_and_labels	naaqs ozone areas plant
30.1.43		-1	referrer_url	(\d+)
30.1.47		-1	document-summary	age a gt
30.1.56		-1	object-title	ets cem
30.1.60		-1	referrer_url	ttp www.epa.gov/enviro.html
30.1.70		-1	document-summary	chemicals
30.1.71		-1	epa_contacts	415 (\d+)
30.1.74		-1	highlighted	co (\d+)
30.1.75		-1	labels	station (\d+)
30.1.77		-1	object-title	station unit
30.1.78		-1	referrer_url	raw data
30.1.79		-1	theurl	raw/data
30.2	Documents linking to problem docment			
30.2.3		1	object-title	ttn
30.2.4		1	webi_keywords	technology
30.2.10		-1	object-title	draft report
30.2.11		-1	theurl	region5/water
30.2.12		-1	theurl	oust
30.2.13		-1	theurl	uic
30.2.14		-1	webi_keywords	underground storage
30.2.15		-1	webi_keywords	agency grants
30.2.16		-1	webi_title	jobs through recycling
30.2.19		-1	tssms	indicate
30.2.20		-1	tssms	werust1
30.2.21		-1	theurl	glrppr.org/hubs
30.2.22		-1	tssms	eg5oh2o
30.2.24		-1	theurl	fedlaws
30.2.25		-1	theurl	workshop_slides

RuleID	Description	Score	Attribute	Values
30.2.26		-1	theurl	presentations
30.2.27		-1	theurl	envindicators/roe
30.2.28		-1	theurl	water/uic/presentations
30.3	Documents linked fromproblem docment			
30.3.3		1	object-title	ttn
30.3.4		1	webi_keywords	technology
30.3.6		-1	img_alt	disclaimer
30.3.7		-1	webi_description	state
30.3.9		-1	tssms	eg5oopa
30.3.10		-1	object-title	draft report
30.3.11		-1	theurl	region5/water
30.3.12		-1	theurl	oust
30.3.13		-1	theurl	uic
30.3.14		-1	webi_keywords	underground storage
30.3.15		-1	webi_keywords	agency grants
30.3.16		-1	webi_title	jobs through recycling
30.3.18		-1	links	tribal
30.3.19		-1	tssms	indicate
30.3.20		-1	tssms	werust1
30.3.21		-1	theurl	www.glrppr.org/hubs
30.3.22		-1	tssms	eg5oh2o
30.3.23		-1	tssms	paoswer
30.3.24		-1	theurl	fedlaws
30.3.25		-1	theurl	workshop_slides
30.3.26		-1	theurl	presentations
30.3.27		-1	theurl	www.epa.gov/envindicators/roe
30.3.28		-1	theurl	water/uic/presentations

RuleID	Description	Score	Attribute	Values
60	System generated rules second crawl			
60.1	Rules applied to problem docment			
60.1.1		2	contact_url	ttn/naaqs/ozone/contactus
60.1.2		2	object-title	epa ttn naaqs
60.1.3		2	document-text	nox co so2
60.1.4		2	document-text	emission home page
60.1.5		2	document-text	epa home privacy
60.1.6		2	document-text	transport of ozone
60.1.7		2	document-text	resources file utilities
60.1.8		2	document-text	page ozone implementation
60.1.9		2	links_and_labels	home page
60.1.10		2	object-title	ttn naaqs
60.1.11		2	document-text	drinking water
60.1.12		2	document-text	scc descriptions
60.1.13		-2	document-text	to prairies
60.1.14		2	document-summary	high-quality scientific
60.1.15		2	contact_url	maia/html/comments
60.1.16		2	labels	sheet
60.1.17		2	webi_title	sheet
60.1.18		2	webi_keywords	sheet
60.1.19		2	object-title	water
60.1.20		2	headings	sheet
60.1.21		2	author	of
60.1.22		2	object-title	water
60.1.23		2	highlighted	totals
60.1.24		2	webi_keywords	brownfields
60.1.25		2	links	and air quality

RuleID		Description	Score	Attribute
60.1.26		2	epa_breadcrumbs	emission trends data
60.1.27		2	links_and_labels	http://www.epa.gov/ttn/naaqs/ozone/areas/index.htm
60.1.28		2	links_and_labels	home page
60.1.29		2	object-title	sheet
60.1.30		-2	object-title	sector resources
60.1.31		-2	object-title	ncee publications regulatory
60.1.32		-2	keyword	tsca valuation value
60.1.33		-2	keyword	legislation market
60.1.34		-2	keyword	control cost cba
60.1.35		-2	document-summary	records are classified
60.1.36	Y	-2	document-summary	office in charge
60.1.37	Y	-2	document-summary	analyses of policies
60.1.38		-2	document-summary	related to epa's
60.1.39	Y	-2	object-title	powerpoint
60.1.40		-2	labels	perfect
60.1.43		-2	keyword	tradeoff trading tsca
60.1.44		-2	keyword	regulation regulatory release
60.1.45		-2	keyword	producer program
60.1.46		-2	keyword	permit
60.1.47		-2	keyword	permit pesticide policies
60.1.48		-2	keyword	omb permit pesticide
60.1.49		-2	keyword	occupational omb permit
60.1.50		-2	keyword	natural occupational omb
60.1.51		-2	keyword	hazardous health human
60.1.52		-2	keyword	equity estimation evaluation
60.1.53		-2	document-summary	protection agency's
60.1.54		-2	document-summary	prevention roundtable
60.1.57		-2	referrer_url	//cfpub.epa.gov/clearinghouse/index.cfm?topicid=c10
60.1.58		-2	links	economic analyses

RuleID		Description	Score		Attribute	Values
60.1.59	Y			-2	author	printing office
60.1.60				-2	subject	pages
60.1.61				-2	referrer_url	//www.epa.gov/imr/download/user/
60.1.62				-2	object-title	register
60.1.64				-2	document-summary	assistance
60.2		Documents linking to problem docment				
60.2.1				2	document-text	foia grants/procurement laboratory
60.2.2				2	document-text	agriculture brownfields cleanup
60.2.3				2	document-text	topics regional adminstrator
60.2.4				2	document-text	and workshops maps
60.2.5				2	document-text	amp development u.s
60.2.6				2	document-summary	and assessment initiative
60.2.7				-2	document-text	public notices
60.2.8				-2	contact_url	region5/water/r5water_comments
60.2.9				-2	document-text	injection control regulations
60.2.10				-2	links_and_labels	notices announcements
60.2.11				-2	document-text	topics other local
60.2.12				-2	document-text	landscape ecology environmental
60.2.13				-2	document-text	satisfaction survey uic
60.2.15				-2	document-text	financing business assistance
60.2.16				-2	document-text	through recycling
60.2.17				-2	document-text	facilities mines_count mines
60.2.19	Y			-2	document-text	hub
60.2.20	Y			-2	document-text	by keyword table
60.3		Documents linked fromproblem docment				
60.3.1				2	document-text	office of wetlands
60.3.2	Y			2	document-text	since the original publication

RuleID	Description	Score	Attribute	Values
60.3.3		2	document-text	is entirely drawn
60.3.4		2	webi_keywords	training and certification
60.3.5		2	webi_keywords	of watershed training
60.3.6		2	document-text	research amp development
60.3.7		2	document-text	gt icr gt
60.3.8		-2	document-text	prevention resource exchange
60.3.9		2	headings	homepage epa home
60.3.10		2	webi_title	of watershed training
60.3.11		2	webi_keywords	of watershed training
60.3.12		2	document-summary	scientific information on
60.3.14		2	subject	emission inventory conference
60.3.15		2	webi_keywords	and certification/ ecosystems
60.3.16		2	img_alt	of watershed training
60.3.17		2	links_and_labels	www.epa.gov/ow/search.html
60.3.19		2	referrer_url	www.epa.gov/ttn/chief/conference/ei13/index.html
60.3.20		2	referrer_url	www.epa.gov/ne/npdes/mirantkendall/index.html
60.3.22		2	webi_keywords	training and certification/
60.3.23		2	object-title	envirofacts warehouse icr
60.3.24		2	contact_url	enviro/html/ef_feedback
60.3.25		2	webi_keywords	education training
60.3.26		2	object-title	envirofacts warehouse
60.3.27		-2	document-summary	great lakes regional
60.3.28		-2	img_alt	pollution prevention roundtable
60.3.29		-2	object-title	management for schools
60.3.30		-2	theurl	hubs/keyword_search