IDSO BEST PRACTICES
Web Crawling

Document # : IDSO-WC-BP-001 Version : 1.0 Effective Date : Draft

APPENDIX 2: CHECKLIST FOR ASSESSING RISK & COMPLIANCE OF WEB HARVESTING.

	WEB HARVESTING REGULATIONS & COMPLIANCE	SECTION
1	All datasets should be obtained by legal means.	
2	Alternative Data purchased or owned by the organization should be assessed for web harvesting compliance.	
3	There should be a documented intent of use for each website that will be Web Harvested.	
4	The organization should create a compliance working group to manage web harvesting within the organization.	
5	The compliance working group should insure that the compliance strategy conforms to the appropriate statutes and laws.	
	WEBSITE ASSESSMENT	SECTION
6	A data collector should assess a website according to the terms of its robots.txt.	
7	A data collector should access websites in a way that the access does not interfere with or impose an undue burden on their operation.	
8	A data collector should not access, download or transmit non-public website data.	
9	A data collector should not circumvent logins or other assess control restrictions such as captcha.	
10	A data collector should not utilize IP masking or rotation to avoid website restrictions.	
11	A data collector should respect valid cease and desist notices and the website's right to govern the terms of access to the website and data.	
12	A data collector should respect all copyright and trademark ownership and not act so as to obscure or delete copyright management information.	
13	A data collector should respect all contact terms with data provider formed by actual, human observed notice, and explicit agreement.	
14	A data collector should not act so as to significantly impair or compete with the business of the website.	
15	A data collector should make the best commercial efforts to preserve an image of the website, then current notification of terms and conditions, and robots.txt file with capture.	
16	A data collector should not generate data in violation of the CAN-SPAM act.	
17	A data collector should seek to use a website's official API when possible.	

18	A data collector should not engage in any criminal or intentionally deceptive conduct or seek to exceed their access authorization.

RISK MANAGEMENT		**SECTION**
19	The organization should have a risk assessment process for identifying the impact level of web harvesting.	
20	A risk assessment should be performed for each website to identify web harvesting impact level.	

WEB HARVESTING POLICIES & MANAGEMENT		**SECTION**
21	The organization should create at least one written procedure for web harvesting management.	
22	The definition of web harvesting should be defined in the procedure.	
23	A synopsis of all relevant privacy laws, regulations, and policies should be included in the procedure.	
24	The departments and individual roles and responsibilities for web harvesting should be defined.	
25	The roles, responsibilities, and response for web crawling incidents should be defined.	
26	The websites that will be crawled should be assessed to mitigate risk.	
27	Guidance should be provided on restrictions of data collection, disclosure, sharing, storage and use of web harvesting within the organization.	
28	The procedure should specify an interval for review of Web Harvested websites and frequencies to determine whether the web harvesting is relevant and necessary.	
29	The procedure should include instructions for handling a security or privacy breach as a result of web harvesting.	

WEB HARVESTING REVIEW AT REGULAR INTERVALS AND AT LEAST ANNUALLY		**SECTION**
30	A review of Web Harvested websites and crawling frequencies should be conducted no less than once per year to determine whether the web harvesting of the particular website(s) is relevant and necessary.	
31	The organization should update their Risk Assessment document at least annually.	
32	Organizations should regularly review audit records of inappropriate or unusual activity surrounding web harvesting processes.	

IDSO BEST PRACTICES
Web Crawling

Document # : IDSO-WC-BP-001 Version : 1.0 Effective Date : Draft

33	The web harvesting procedure should be updated at least annually.

WEB HARVESTING INCIDENTS		SECTION
34	A response plan to handle web harvesting incidents should be detailed in the web harvesting procedure.	
35	Information obtained through detection, analysis, containment, and recovery should be collected to help protect against future incidents.	

TRAINING		SECTION
36	The organization should provide training on web harvesting control policies including responsibilities, implementation, and access controls.	
37	The organization should provide incident response training for web harvesting.	