IDSO BEST PRACTICES
Web Crawling

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

1. PURPOSE

The purpose of this document is to provide guidance, processes, and risk management for web harvesting. The document includes considerations for best practices and compliance with the state, regional, and federal laws when using Alternative Data extracted or harvested from the web.

2. SCOPE

This document includes guidance, risk assessment, and management of Alternative Data obtained through web harvesting (i.e., web crawling or web extraction techniques).

3. AUDIENCE

The primary audience are managers and compliance teams in an investment company interested in best practices, procedures, and regulatory guidance related to web harvesting. This document applies to teams within the following company profiles that participate in the Alternative Data ecosystem:

3.1 **Raw Data Originators:** Companies that collect the data for the sole purpose of selling or generating data as part of their core business.

3.2 **Research Providers:** Companies that use raw data to produce original research and derived signals. Research Providers occasionally generate data through surveys, measurements, and other means.

3.3 **Aggregators:** Companies that aggregate and enrich data with the intent to sell, license or distribute data to the investment community.

3.4 **Consumers:** Investment professionals and companies that use data to add value to their investment process.

4. TABLE OF CONTENTS

Section ***Page***

5. BACKGROUND

Web harvesting is a widely-used technique for producing Alternative datasets for investment research on both the buy and sell side. Although operating a web crawler is a legal and commonly conducted activity on the internet, firms need to manage their risk associated with the practice.

5.1 DEFINITION

Web harvesting, also termed web scraping or web extraction is the practice of extracting data from websites. Web scraping can be done manually or automatically via custom code, web scraping software, bots, or web crawlers. The web scraper copies content from web pages to a central database or spreadsheet for data processing, aggregating and analysis.

5.2 RESPONSIBILITY

All entities that use web-harvested Alternative Data should adhere to the legal obligations of the website, contractual and fiduciary obligations between organizations, and state, regional, and federal statutes and regulations. The regulations require businesses to take reasonable security measures to protect both individuals and organizations. Although it may take years for well-defined regulations to come to fruition, internal standard operating procedures should be developed and adhered to in the interim based on current knowledge of the regulatory landscape. The use of Alternative Data obtained through unethical or illegal means can result in legal harm to individuals and organizations. Therefore, data obtained via web extraction needs to be accessed and used per the appropriate laws. Also, each organization's internal procedures should capture business risks in conjunction with the appropriate legal and compliance counsel.

5.3 BEST PRACTICES FOR WEB HARVESTING

The purpose of this document is to provide guidance in the compliance, risk management, and procedures for web harvesting in the Alternative Data segment of the financial industry. Each document section provides guidance and best practices for accessing and managing Web Harvested data. The appendices provide a risk assessment template and checklist for the best practices. By using these guidelines, organizations who use web harvesting procedures will avoid losses and legal action associated with inadequate web harvesting procedures and compliance.

6. INTRODUCTION

Web harvesting, or crawling or scraping, is used to extract information from websites. It was created to copy websites to generate a web page index for search engines. Over the years,

many other purposes for web crawling have developed. Other objectives include summarizing, aggregating, indexing, or extracting information for a third-party organization's internal purposes. Web harvesting has become a popular practice and has become widely used for the compilation of Alternative data for the financial industry. Along with the standard web crawling processes, there are cases where web crawling has been used for unethical or illegal practices. Some of these improper practices include:

- A web crawler making unauthorized copies of copyrighted information on a website
- Website access that circumvents technical restrictions
- Website access that impairs the performance of the website
- Obtaining information to gain a competitive edge over that website
- A breach of terms stated on the website

Investment Managers must comply with contractual and fiduciary obligations and federal and state regulations while extracting data from the web or purchasing web-harvested datasets. The regulations essentially require that the organization:

- Provide reasonable security for data, systems, and communications
- Disclose breaches to affected parties and regulators, and disclose material risks

The majority of laws are concerned with protecting personally identifiable information (PII) (e.g., social security numbers or home addresses). Compliance and management of PII are detailed in Doc. No. IDSO-PII-BP-001. There are federal, state and regional laws that apply to organizations that perform web crawling and for keeping data secure. These laws essentially ask organizations to "take reasonable security measures." The goal of organizations that use Alternative data is to understand the markets better; therefore, these organizations are typically not looking to re-distribute the data, compete with the host website, or put a significant burden on web services.

Most of the laws and case history related to web crawling does not directly apply to the investment industry. However, the information that currently exists can help the Alternative data industry proactively comply and adhere to best practices to avoid direct or indirect damage through opportunity or cost to data organizations.

7. REGULATIONS ON WEB HARVESTING AND FAIR INFORMATION PRACTICES

There are many Federal and State statutes and regulations such as the widely-recognized Digital Millennium Copyright Act (DMCA), the Computer Fraud and Abuse Act ("CFAA"), and the Trespass to Chattels. These laws are referenced in the case history[1] involving web

[1] Legal Cases Relating to Web Harvesting. Eagle Alpha, January 27, 2016. https://s3-eu-west-1.amazonaws.com/ea-documents/papers/Legal+Cases+Relating+to+Web+Harvesting_PDF.pdf Last Accessed September 28, 2017.

crawling. This document (Doc. No. IDSO-WC-BP-001) uses the web crawling case history[1] to define the best practices for web harvesting.

Each organization should consider its legal obligations and determine its web harvesting risk level. Decisions about a law or statute may require consultation since applicable legislation and regulations change over time. Preventing web harvesting malpractice requires knowledge of regulations and commitment towards applying best practices.

7.1 DIGITAL MILLENNIUM COPYRIGHT ACT (DMCA)

The Digital Millennium Copyright Act ("DMCA") has been used against the owners of web crawlers that circumvent technical restrictions to access copyrighted web content. The DMCA states that the website owner should have a valid copyright with a technical barrier to prevent users from accessing certain content. Web crawlers that circumvent technical restrictions on websites and then copy and distribute information are in violation of this act. For more information about this law, visit the U.S. Copyright Office.[2]

7.2 COMPUTER FRAUD AND ABUSE ACT (CFAA)

The Computer Fraud and Abuse Act ("CFAA") is a federal anti-hacking statute. Some cases have used CFAA to claim that the web crawler has accessed information from a protected computer resulting in losses of at least $5000 in value. The definition of "loss" includes the cost of responding to a network intrusion. Several websites have successfully used CFAA against web crawlers because they were forced to spend money to block unauthorized access. For more information about this act, visit justice.gov.[3]

7.3 TRESPASS TO CHATTELS

Some cases involving automated web crawlers have succeeded on a theory of trespass to chattels. Trespass to Chattels is a civil wrong, whereby one party intentionally interferes with another person's lawful possession of a chattel (movable personal property). The interference causes a decrease in the condition, quality or value of the personal property. For more information on Trespass to Chattels, download *Torts: Cases and Context*[4] Volume 2 Chapter 21 by E. E. Johnson.

8. WEB HARVESTING GUIDELINES

When considering the web harvesting of particular websites, a legal review and risk assessment of each website should be conducted. Implementing the proper oversight and procedures within the organization that creates the web crawler ("creator") will ensure that unauthorized websites do not get crawled. Website files such as the robots.txt and the Terms of Use (TOU) should be

[2] The Digital Millennium Copyright Act of 1998. U.S. Copyright Office, December 1998. https://www.copyright.gov/legislation/dmca.pdf Last Accessed September 28, 2017.

3 Jarrett, Marshall H. and M. W. Bailie. Prosecuting Computer Crimes: Computer Crime and Intellectual Property Section Criminal Division. Published by Office of Legal Education Executive Office for United States Attorneys.

4 Johnson, E. E. Torts: Cases and Context. Volume 2. eLangdell Press, 2016. Chapter 21. Tresspass to Chattels and Conversion. pp. 227 – 275. https://www.cali.org/books/torts-cases-and-contexts-volume-2# Last Accessed September 29, 2017.

reviewed to determine access limitations on the website. web harvesting is low risk as long as non-copyrighted content is extracted, the website access does not affect website usage, and the content is used for internal research and development. When implementing internal procedures or purchasing crawled data ("buyer"), the organization should verify that:

- The crawler complies with the *Terms of Use (TOU)* and the *ROBOT.TXT* instructions.
- The crawler does not overload the website's servers or affect the *website operation.*
- The crawler extracts only factual information and not *Copyrighted* content.
- The crawler accesses *publicly available information* (the crawler did not access any non-public areas of the website.
- If an *API* is available, it is used instead of web scrapping.
- The crawler does not extract information for *competitive purposes* with the website host.

A brief description of these guidelines is detailed in this section. Using an external party for building and maintaining the crawlers may provide an extra level of protection from legal claims because the claims are usually filed against the company that operates the crawler. However, certain organizations decide not to take this route because they do not want to disclose the web resources being collected or lose control over their compliance procedures.

8.1 TERMS OF USE (TOU)

The ability to purchase products and enter into contracts on the internet evolved so quickly that the regulations did not keep up. Websites often use one of two types of agreements for purchases and Terms of Use (TOU) for software: Browse-wrap and Clickwrap agreements. The Browse-wrap agreement can be accessed on a website through a link to another page that contains the terms and conditions. With the browse-wrap agreement, there is usually no affirmation that the customer has viewed the terms on the linked web page. In contrast, the Clickwrap Agreement requires the customer to review the terms of the agreement through a series of pop-up windows. Therefore, the Terms of Use (TOU) of each website being harvested should be evaluated, and the assessment should be from a Browserwrap versus a Clickwrap perspective.

Courts have been hesitant to enforce Browser-wrap agreements while allowing the enforcement of Clickwrap Agreements. Courts believe that users are more likely to be informed of the terms and conditions when they are required to accept terms and conditions by confirming through a mouse click or series of actions. The courts have ruled that a crawler is not bound to the TOU when the website has used a Browserwrap agreement.

8.2 ROBOTS.TXT

After the development of the internet search engine, the "Robots Exclusion Protocol" was established to allow a website to provide instructions to web crawlers through the use of "robots.txt" files. These files specify which crawlers are allowed to access the website and the pages on the website that the robots are allowed to crawl. The three major parameters used in

robots.txt files are “User-agent,” “Allow,” and “Disallow.” The “User-agent” parameter specifies a particular web-crawler, “Allow” tells the crawler what directories it is allowed to crawl, and “Disallow” tells a crawler directories that it cannot crawl. The robots.txt files only provides instructions for web crawlers; these files cannot prevent the robot’s access.

8.3 WEBSITE OPERATION

Web harvesting should occur so that it does not interfere with the website or impose an undue burden on the website’s operation. While evaluating a website, the appropriate level of requests over a certain time interval should be specified along with the appropriate method of measuring the total monthly traffic load. The number of requests should be kept to a level that does not interfere with the website’s operation (i.e., below 0.5% of the total traffic). A geographic HTTP request distribution can be used to help minimize the load on a particular server.

8.4 COPYRIGHT

A web crawler should avoid making complete copies of a website. A web crawler that makes a copy of a website that contains copyrighted content for archiving or redistribution may infringe on copyright laws. The content that is copied or extracted should be factual information only, such as pricing, dates, and locations. If non-factual information is copied, it should be stored long enough for analysis and then be discarded.

8.5 PUBLIC INFORMATION

The information that a web crawler extracts from a website should only be public information. The web crawler should not access non-public content and should be designed to detect protected or secured content and not bypass technical barriers to access certain parts of a website.

8.6 APPLICATION PROGRAM INTERFACE (API)

Many websites offer an option to download data using an API. The API may be a free or paid option, but using the API when available is recommended over directly scraping the website. If an API is available, and not utilized, some website hosts may view this as lost revenue. Also, the web crawler could potentially be accessing data that the host did not intend to distribute.

8.7 COMPETITION

A web crawler should not extract information from a website to gain a competitive edge with the host website.

9. RISK MANAGEMENT

Risk management is the process of identifying and managing risks to individuals and organizational operations. A comprehensive risk evaluation should be performed before the commencement of a significant web harvesting project. The risk assessment process should be incorporated into an internal compliance procedure(s), and should help to determine if a website should be crawled by rating important factors or through assigning an “impact level.” A methodology for determining risk and a compliance checklist is provided in Appendices 1 and 2.

9.1 IMPACT LEVEL

'Impact' can be defined as the magnitude of harm that could potentially result from the misuse of web harvesting. The impact level is based on the implementation of the guidelines mentioned in Section 8. The web harvesting impact level can be classified into three categories based on the potential misuse of web crawling as shown in Table 9-1. An example of a risk assessment that uses the categories of high, moderate, and low is provided in Appendix 1. Factors used to identify impact level in Appendix 1 are website terms, website operation, public information, copyright, and competition.

Table 9-1. Web Harvesting Impact Level[4]

IMPACT LEVEL	DESCRIPTION	CONSEQUENCES
HIGH	Disregarding website terms, copyrights, secured information, or competing with the host organization could have a **severe or catastrophic effect** on an organization or individuals.	The organization may not be able to perform one or more of its primary functions. There may be significant damage to organizational assets. There may be significant financial loss.
MODERATE	Disregarding website terms, copyrights, secured information, or competing with the host organization could have a **severe adverse effect** on an organization or individuals.	The organization may experience a significant degradation in mission capability for a certain extent and duration. There may be damage to organizational assets. There may be financial loss.
LOW	Disregarding website terms, copyrights, secured information, or competing with the host organization could have a **limited adverse effect** on an organization or individuals.	The organization may experience a degradation in mission capability to a certain extent and duration. There may be minor damage to organizational assets. There may be minor financial loss.

10. COMPLAINTS & INCIDENT RESPONSE

Organizations should designate one or more individuals to address incoming complaints from third-party websites. Communication of complaints can occur through email, vendors, and proxy services

and these communication sources should be monitored appropriately to ensure that all complaints are read and handled properly. If a "Cease and Desist" complaint is received, it must be quickly escalated to the appropriate individuals within the organization. Each complaint that is received should be recorded, investigated, and validated. Therefore, the organization should have a quality system for managing complaints. After a complaint is validated, an appropriate response to the complainant should be sent.

10.1 INCIDENT RESPONSE

Incidents involving web harvesting have the potential to damage an organization's reputation and incur substantial costs and time. Organizations who use web harvested Alternative Data should develop policies that describe when and how companies should be notified, when and if misuse should be reported publicly, and whether to provide remedial services to organizations affected. Organizations should integrate such policies into their existing web harvesting procedures. Table 10-1 describes the four phases for handling web crawling incidents: preparation, detection and analysis, and post-incident recovery.

Table 10-1. Phases for Handling Web Crawling Incidents[5]

INCIDENT PHASE	DESCRIPTION
PREPARATION	Organizations should create response plan for web harvesting and incorporate the plans into an existing procedure. The policies and procedures should be conveyed to the organization's staff through training and awareness programs.
DETECTION AND ANALYSIS	Complaints or internal quality systems for detection may be useful for web harvesting incidents. However, additional procedures for incident handling may be necessary to insure successful handling and recovery.
POST-INCIDENT ACTIVITY	Information obtained through detection, analysis and recovery should be collected for sharing within the organization to help protect against future incidents.

11. WEB HARVESTING POLICIES & MANAGEMENT

Web harvesting compliance and processes should be clearly documented in explicit policies and communicated through training and awareness. Internal procedures help an organization ensure that best practices are being followed. Some of the topics that the procedure should specify are:

- The process for each web harvesting project

[5] Security Controls are from National Institute of Standards and Technology Special Publication 800-122 *Guide to Protecting the Confidentiality of Web Harvesting,* April 2010.

- A risk assessment process and template
- A compliance strategy that conforms to the appropriate statutes and laws
- Web crawling project approval by the appropriate organization executive or manager
- Individuals accountable for implementation, functioning, and compliance of the appropriate controls in the organization
- Reviews of the company's procedures to remain in compliance

Table 11-1 lists topics that should be included in one or more procedural document(s) for compliance, control, and incident response of web harvesting for companies who use Alternative Data. The document(s) should be reviewed annually and should include senior management.

Table 11-1. Topics to Include for Written Web Harvesting Policies

DOCUMENT SECTION	DESCRIPTION
DEFINITION	A formal definition of web harvesting (scraping/crawling) should be included in the procedure.
APPLICABLE PRIVACY LAWS, REGULATIONS, AND POLICIES	A synopsis of all relevant privacy laws, regulations, and policies should be included in the procedure.
ROLES AND RESPONSIBILITIES	The departments and/or individual roles and responsibilities for using web harvesting should be defined. The roles and responsibilities in responding to web harvesting-related incidents and reporting should also be defined.
ACCESS RULES	The access rules for web harvesting must be carefully considered. The organization should conduct a periodic review of personnel handling web harvesting.
RISK ASSESSMENT	The organization should assess risk associated with web harvesting. The risk assessment process should be reviewed at least annually, and include senior management responsible for managing web harvesting.
INCIDENT RESPONSE AND DATA BREACH	A response plan to handle web harvesting incidents should be detailed in the procedures. The response plan should address communications with all relevant individuals and include response policies for misuse.
AUDITS	The organization should develop a policy of auditing systems that control web harvesting.

11.3 COOPERATION & COLLABORATION

Organizations who use Alternative Data should promote close cooperation among senior management and legal counsel when addressing issues related to web harvesting. Cooperation and collaboration of the relevant internal and external experts help to prevent the misuse of web harvesting by strictly adhering to internal policies and procedures and adequately training staff to oblige by these systems and procedures.

12. EDUCATION, TRAINING, AWARENESS

After policies and procedures have been formalized, training and education is essential to a successful web harvesting program. Laws and regulations may specifically require training for staff, managers, and contractors. An organization should have a training plan and implementation approach, and an organization's leadership should communicate the web harvesting policies to its staff. The goal of training is to build knowledge and skills that will enable staff to use best practices when working with web harvesting. In addition, training updates that include compliance training for all employees is required. Table 12-1 describes training of website access, web harvesting training, and incident response. For additional information on developing a training program, refer to NIST SP 800-50, Building an Information Technology Security Awareness and Training Program.[6]

Table 12-1. Training for Web Harvesting

SECURITY CONTROL	DESCRIPTION
WEBSITE ACCESS	The organization should provide training on website access including responsibilities, implementation, and controls.
WEB HARVESTING TRAINING	The organization should provide web harvesting training to system users as (1) part of the initial training and as (2) procedures are updated. The organization should document training activities and retain individual training records.
INCIDENT RESPONSE TRAINING	The organization should provide incident response training for users involved with web harvesting.

[6] National Institute of Standards and Technology Special Publication 800-50, Building an Information Technology Security Awareness and Training Program, October 2003.

13. REFERENCES/RELATED PROCEDURES

National Institute of Standards and Technology Special Publication 800-122 *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* April 2010.

National Institute of Standards and Technology Special Publication 800-39 *Guide to Managing Information Security Risk,* March 2011.

National Institute of Standards and Technology Special Publication 800-30 *Guide to Conducting Risk Assessments,* September 2012.

The Digital Millennium Copyright Act of 1998. U.S. Copyright Office, December 1998. https://www.copyright.gov/legislation/dmca.pdf Last Accessed September 28, 2017.

Legal Cases Relating to Web Harvesting. Eagle Alpha, January 27, 2016. https://s3-eu-west-1.amazonaws.com/ea-documents/papers/Legal+Cases+Relating+to+Web+Harvesting_PDF.pdf Last Accessed September 28, 2017.

Johnson, E. E. Torts: Cases and Context. Volume 2. eLangdell Press, 2016. Chapter 21. Trespass to Chattels and Conversion. pp. 227 – 275. https://www.cali.org/books/torts-cases-and-contexts-volume-2# Last Accessed September 29, 2017.

14. APPENDICES

14.1 Appendix 1: Web Harvesting Risk Assessment Example.

14.2 Appendix 2: Checklist for Assessing the Risk & Compliance of Web Harvesting for Alternative Data Use.

15. REVISION HISTORY

The following revision history contains the changes since the last revision of the document.

Version Number	*Change Description*	*Revised By*	*Date of Revision*
1.0	New procedure		

IDSO BEST PRACTICES
Web Crawling

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

Signature Page

Document Author:	**IDSO Designee:**
Written By:	
Date:	
Document Approver:	**IDSO Management:**
Approved By:	
Date:	

IDSO BEST PRACTICES
Web Crawling

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

APPENDIX 1: WEB HARVESTING RISK ASSESSMENT EXAMPLE

The objective of this risk assessment is to determine the impact level of web harvesting to determine whether to crawl a website. There are three impact levels: low, moderate and high. Factors used to identify impact level are website terms, website operation, public information, copyright and competition. Table A1-1 summarizes the criteria to assign the appropriate level for each factor.

Table A1-1. Factors Used to Identify Impact Level

	LEVEL 1	LEVEL 2	LEVEL 3
WEBSITE TERMS	The Terms of Use (TOU), robots.txt, and other actual, explicit, or observed terms have not been evaluated.	Some of the website terms have been evaluated.	The Terms of Use (TOU), robots.txt, and other actual, explicit, or observed terms have been evaluated and the website can be accessed.
WEBSITE OPERATION	The data is accessed in a manner that interferes, impairs or burdens the website operation.	Access of the data may interfere or burden the website during peak operating hours.	The data is accessed in a manner that does not interfere, impair or burden the website operation.
PUBLIC INFORMATION	The extracted data consists of non-public information only.	The extracted data mostly consists of public information.	The extracted data consists of public information only.
COPYRIGHT	Information that is copyrighted or trademarked may have been extracted.	Information that is copyrighted or trademarked may have been extracted.	No copyrights or trademarks have been violated during the web extraction operation.
COMPETITION	The collected data is used to compete with the business of the host website.	The collected data may be used to compete with the business of the host website.	The collected data is not used to compete with the business of the host website.

Table A1-2. Example Risk Assessment to Identify Impact Level

WEB HARVESTING DATA FIELD	WEBSITE TERMS	WEBSITE OPERATION	PUBLIC INFORMATION	COPYRIGHT	COMPETITION	SUM
WEBSITE #1	3	3	3	3	3	15
WEBSITE #2	2	2	3	3	3	13
WEBSITE #3	3	3	2	3	3	14
WEBSITE #4	3	1	3	1	3	11
WEBSITE #5	3	3	3	3	3	15

There are many ways to perform a risk assessment and the specific method used can be determined by the organization. Table A1-2 shows an example risk assessment using the factors summarized in Table A1-1. The steps for filling out this table are:

1. List all the websites in the first column.
2. After all the websites have been listed, assign a level of 1, 2, or 3.
3. Assign a level for “Website Terms”.
4. Assign a level for “Website Operation”.
5. Assign a level for “Public Information”.
6. Assign a level for “Copyright”.
7. Assign a level for “Competition”.
8. After the level for each factor has been assigned, sum the total for the numbers in the row in the last column.
9. Using the sum, the ‘Impact Level’ can be assigned using Table A1-3.

Table A1-3. Security Level

SCORE	IMPACT LEVEL
< = 12	High
BETWEEN 12 & 14	Moderate
BETWEEN 14 & 15	Low

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.	