

Improving the Timeliness of Environmental Management Information Systems with Data Crawling Techniques

Jay Rajasekera

Graduate School of International Management
International University of Japan

Maung Maung Thant

Graduate School of International Management
International University of Japan

Ohnmar Htun

Nagaoka University of Technology

June 2009

Improving the Timeliness of Environmental Management Information Systems with Data Crawling Techniques

Jay Rajasekera*, Maung Maung Thant*, Ohnmar Htun**

*International University of Japan

**Nagaoka University of Technology

Contact e-mail: jrr@iuj.ac.jp

Abstract: With global warming taking center stage, it is becoming clear that environmental information plays a critical role for monitoring, educating, and taking control measures. Currently the environmental data are gathered in a somewhat hierarchical system where mostly governments, NPOs and other organizations collect the data and feed in to world organizations for final analysis and monitoring purposes causing considerable time lags. Using crawling methods and accessing data stored in multiple data sources, and storing in an appropriately designed database could be a key to monitor the global warming in a much more comprehensive and timely manner. An automated software agent can be designed to maintain a collection of data sources. Such Environmental Management Information System (EMIS) works like a centralized system extracting data in real time from websites, which can serve as the springboard for various tasks. The objective of this paper is to present a review of existing EMISs and discuss the practicability of using data crawling techniques in EMIS.

Key Words: Environmental Management Information Systems, Internet, Data Crawling, Databases

1. Introduction

The world is waking up to the realities of environment management and global warming. While world bodies, governments, large corporations, and the concerned public are talking about what to do to stop the global warming, it is becoming clear that information about global warming and the public awareness play a critical role.

The sources that affect environment and the global warming are distributed around the world. Addressing the global environmental changes, without comprehensively accounting for the effects from such sources simply has no meaning. But, the environment related data from such sources are in all kinds of forms residing at millions of locations around the world and are often published on corporate, NGO, and governmental websites. Can the rapidly developing Internet crawling technologies, that are being used to monitor and govern the Internet be used to collect such environmental data from around the world and process accordingly to monitor and govern the environment is the main theme of the proposed research plan.

Currently the environmental data are gathered in a somewhat hierarchical system where mostly governmental NPOs collect the data and feed in to world organizations for final analysis and monitoring purposes (Refs: CAIT & UNFCCC1). Inherent in this hierarchical process are reliability, timeliness, and compatibility issues.

In May 2009, the authors examined the websites maintained by UN and other organizations, which are in the forefront of addressing environmental concerns, for the timeliness of the data. The results, as seen in Table 1 clearly show that there is a significant time lag of updating the data.

Description of data	Maintained by	Time Span (as of May 2009)
Greenhouse Gas Inventory Data	United Nations Framework Convention on Climate Change (UNFCCC)	1990-2006
Greenhouse Gas Emissions By Country	Carbon Planet	1994-2003
CO ₂ Emission per Capita	UN Statistic Division	2004
Greenhouse Gas Emissions per Capita	Globalis - UNEP Grid	2004
CO ₂ Emission	UN Statistic Division	1990-2004
CO ₂ Emission	Energy Information Administration	1990-2007
Emissions and Pollution	The World bank	1990-2005
GHG Emission	World Resource Institute: Climate Analysis Indicators Tool (CAIT)	1850-2005
GHG Emission	Carbon Dioxide Information Analysis Center	1751-2005

Table 1: Lack of Timeliness on Reporting Environmental Data (Sources: Respective websites)

Using crawling methods and accessing data stored in multiple data sources, and storing in an appropriately designed database could be a key to monitor the global warming in a much more comprehensive and timely way than currently being done. An automated software agent can be designed to maintain a collection of data sources. Such Environmental Management Information System (EMIS) works like a centralized system extracting data from websites, which can serve as the springboard for various tasks, including information retrieval, event monitoring and data comparison in multiple data repositories (Refs: NEC Japan, Toshiba Japan, Pinder and Slack, and Kropp and Scheffran).

Lack of well planned EMISs is one reason for the delay in gathering environmental data in a timely manner. The possibility of using a publicly available domain to organize the data so that they can be collected in appropriately and organized in a suitable database is examined in this paper. Also a prototype framework is discussed.

2. Environment Management Systems

The dictionary definition of “global warming” is generally expressed as “Global warming is increasing in the earth's atmospheric and oceanic temperatures widely predicted to occur due to an increase in the greenhouse effect resulting especially from pollution.”

Today emission of green house gases (e.g. CO₂) is claimed to be causing many environmental effects. What causes CO₂ emissions are many human activities such as high usage of fossil fuels and deforestation. CO₂ is only one factor causing environment changes.

The dangers to the world that can be caused by green house gases had raised concerns worldwide. World bodies such as United Nations (UN), and other organizations, such as NGOs have become quite active in trying to develop measures to control or reduce greenhouse gases (Ref: UNFCCC2). One example of world bodies coming together to address the environment issues was the Kyoto Protocol meeting held at the western Japanese city of Kyoto in 1997 (Ref: UNFCCC3).

An important element of environment control is monitoring by using data related to elements that cause harm. Since we are talking about “global warming,” we need to look at the environment related data from all the sources from anywhere in the world. Today a world body such as UN

collects such data via a more or less hierarchical system depending on UN or governmental sources (Ref: UNFCCC1).

But, in reality, governments or UN are not the creators of greenhouse gases. It is various businesses, factories, farms, transportation systems, utilities, and humans that create most of such gases. Some of such entities generating greenhouse gases report their activities, but many perhaps do not.

Even when they report, often, not all the required data are reported. And there is no uniform format followed by all the entities. The “data pipeline” from the source until data reaches an environmental authority, such as UN, which compiles and report using aggregate methods, can be a long process consuming time. The data in UN site is about one year old (see Table 1).

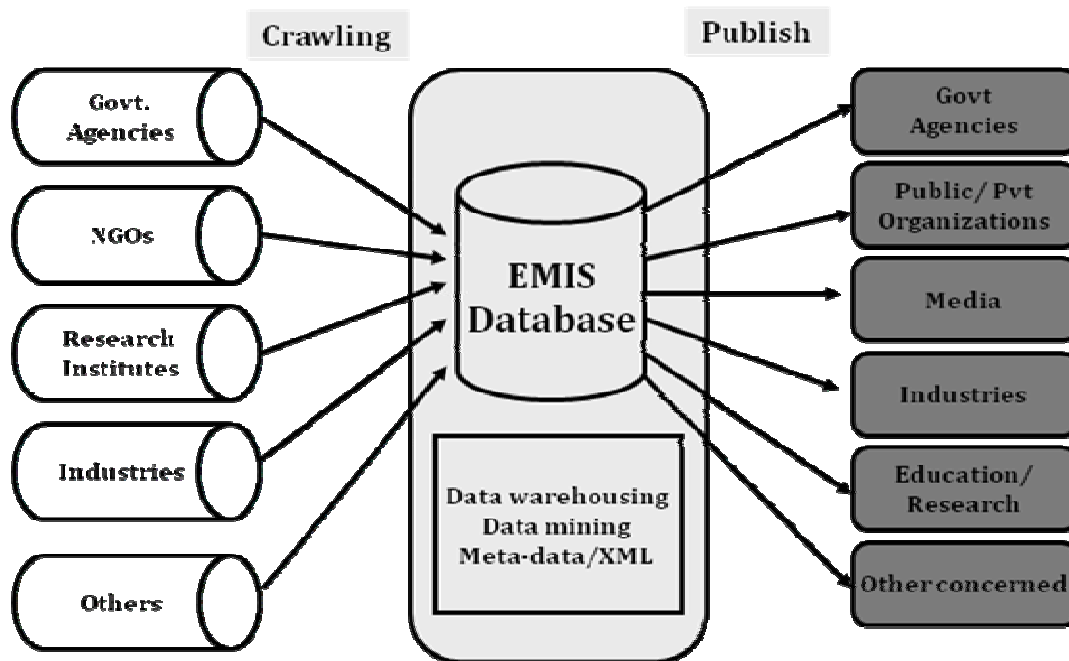


Figure 1: EMIS Process Model

Also, the reporting mechanisms may not necessarily be uniform over all the organizations, though there are some efforts to unify the environmental data. The US EPA had proposed an Environmental Management System (EMS), which is a set of processes and practices that enable an organization to reduce its environmental impacts and increase its operating efficiency (Ref. US EPA). In Japan NEC Corporation and Toshiba had come up with IT-based environment management systems (Refs: NEC Japan, Toshiba Japan).

In addition to what we just mentioned, many research institutes and organizations have developed the various kinds of Environment Management Information Systems (EMIS) to gather data into computer databases and distributing them via Internet.

Indeed Internet is an ideal media for accumulating and distributing environmental data. It is a real-time-two-way communication system which is much more effective and cheap than the other media such as Telephone (Ref. IGES, Globalis).

In order to develop an effective Internet-based EMIS, one must pay attention to collecting data and distributing the compiled data as information. On the data gathering part, one has to pay attention to comprehensiveness, credibility, timeliness, and ease of inputting. On the dissemination side, one has to pay attention to organization of data in order to make it easy to analyze and see the

statistical relationships. And, attention needs to be paid to the kind of audience that data may be targeted. For example data suitable for general public may be different from what needs to be shown to a governmental organization.

Figure 1 shows a schematic view of the data gathering and data decimation process.

3. Crawling for Data

In order to gather data from vast number of sources spread over the Internet, an often used method is called “crawling.” A dictionary definition of crawling is as follows.

“A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner.”

Crawling, sometimes also known as spidering, is widely used for collecting data from Internet websites. Web crawlers are often used by search engines in order to speed up the search procedures by indexing already visited pages (Ref: Web Crawler).

Web crawler is indeed a software agent, sometimes called “bot,” which can work with a set of pre-specified URLs called “seeds.” As the crawler visits each seed, it collects data about all the hyperlinks in that seed and makes something called “crawl frontier.” This crawl frontier, which is basically a list of URLs, can then be used to gather further data.

The typical high-level architecture of a standard Web crawler is shown in Figure 2. It has a scheduler and multi-threaded downloader together with URL queries and data storage.

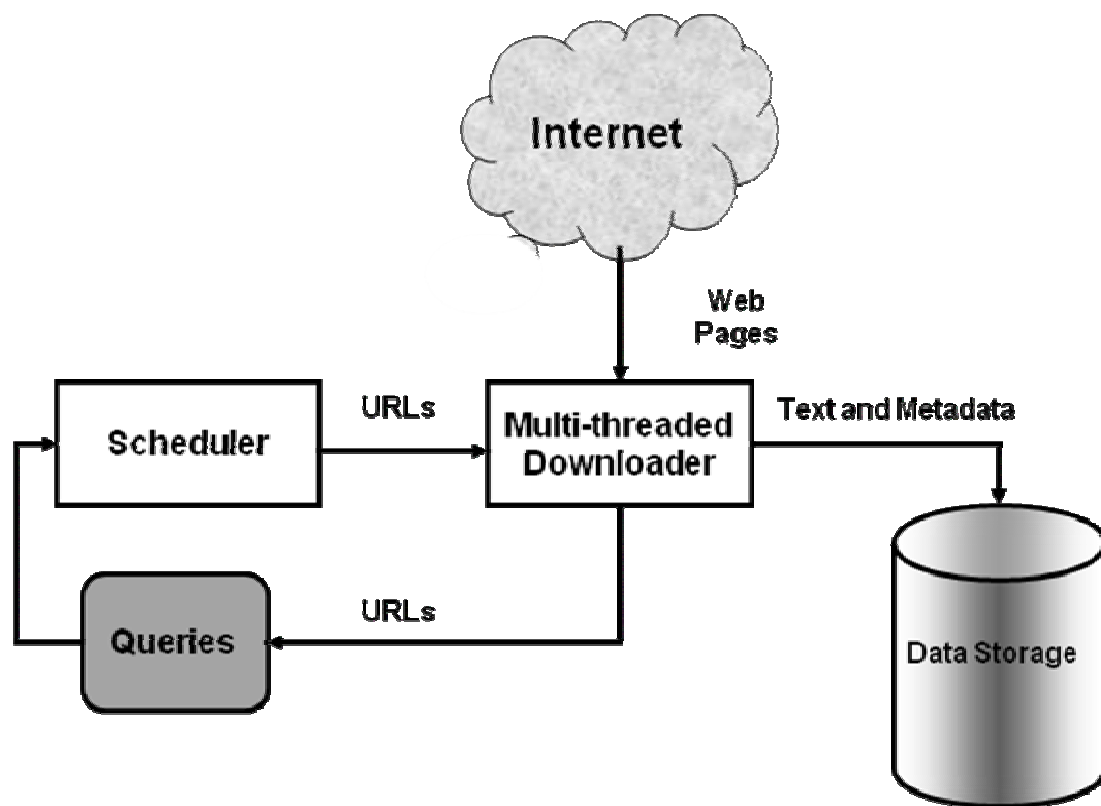


Figure 2: High-level architecture of a Web crawler

Some of the examples of published Web crawler architectures for general purpose crawling are RBSE, WebCrawler, World Wide Web Worm, Internet Archive Crawler, Google Crawler, WebSphinx, CobWeb, Mercator, WebFountain, PolyBot, WebRace, Ubicrawler and FAST Crawler.

Moreover, there are varieties of general purpose web crawlers with different components and features, written in many different programming languages such as Java, C, C++, Perl, Python, PHP and so on. Some of the general purpose Open Source license web crawlers and their features are listed in the Table 2.

Crawler Name	Language	License	Database Support
Nutch	Java	Apache	No
Heritrix	Java	GPL	No
WebSphinx	Java	Apache	No
HTTrack	C	GPL	No
Sphider	PHP	GPL	MySQL, SQLite
Sphinx	PHP	GPL	MySQL, PostgreSQL

Table 2: Open Source license web crawlers

4. Proposed Crawling Method

According to the Netcraft survey in May 2009, there are 235,890,526 websites¹ discovered by their automatic web-exploring "spider" software and it is estimated that there were more than 30 billion of web pages under the World Wide Web (WWW). Search engines are tools to extract information from such a large number of web pages. Generally, search engines visit the sites using automatic programs called crawler or spider, analyze the documents and index the billion of pages. The search engine operations of web crawling, indexing and searching enable the users to query information, typically using keywords (See Figure 3).

However, most of the crawling techniques using by search engines are designed for general purpose, so the search results from search engine could be irrelevant and difficult to organize for comprehensive understanding of a particular subject or issue such as global warming. In addition to this, most of the search engines cannot index the information from the invisible web which includes unlinked pages, excluded pages and user-protected databases.

It could be a solution if the environmental data can be collected by using an intelligent software agent, which can automatically retrieve information, monitor the events, analyze and corroborate the data from multiple web sites, and database together.

¹ Source: http://news.netcraft.com/archives/web_server_survey.html

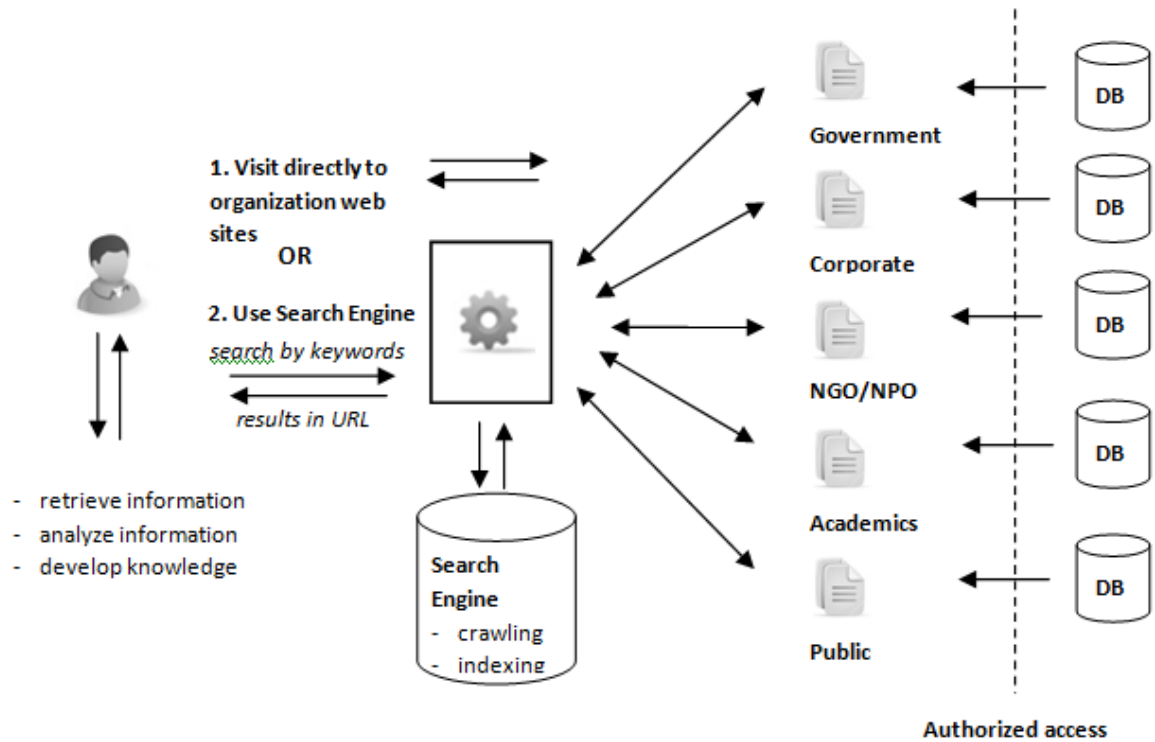


Figure 3: Information retrieval using general purpose search engine

The proposed conceptual techniques in this solution involve the following steps:

1. Retrieving information from the Web by using a focused crawler or a customized search engine which can collect environmental related data
2. Storing the search results in the appropriately designed EMIS database and updates the results eventually
3. Analyzing and data mining the results stored in the EMIS database for further knowledge and comprehensive solutions development
4. Publishing the information and sharing knowledge by integrating the data mining results

First, an automated software agent feeds the keywords from the EMIS database to the Web crawler. The Web crawler works the process of crawling and indexing based on seed URLs provided by the software agent and return the results. Then, the automated software agent analyzes the keywords, search results and store relevant links in the EMIS database. After information retrieval process, the intelligent engine of the software agent analyzes and data mine the results and integrate information for knowledge discovery. Finally, these results can be published for comprehensive understanding of issues and knowledge sharing (See Figure 4).

5. Crawling – an Evaluation

In the proposed crawling method, there are two important issues relating to the Web crawling and information retrieval. First concern is how to build the focused Web crawler or customized search engine and second is how the intelligent software agent can analyze and extract the specific data.

In the development of a focused Web crawler or customized search engine, the first thing we need to build is a list of web sites (seed URLs) to crawl the environmental data which we are looking for. Here, the question is how we know the list of seed URLs. One possibility is to use the list of environmental related URLs from existing Open Directory Project like DMOZ (Ref: ODP). Second thing we have to consider here is the keyword table. It would be effective if we could build keyword dictionary which is semantically related to the information we are looking for. One more important thing we need to take into account is how to define the rank and relevancy of the search results.

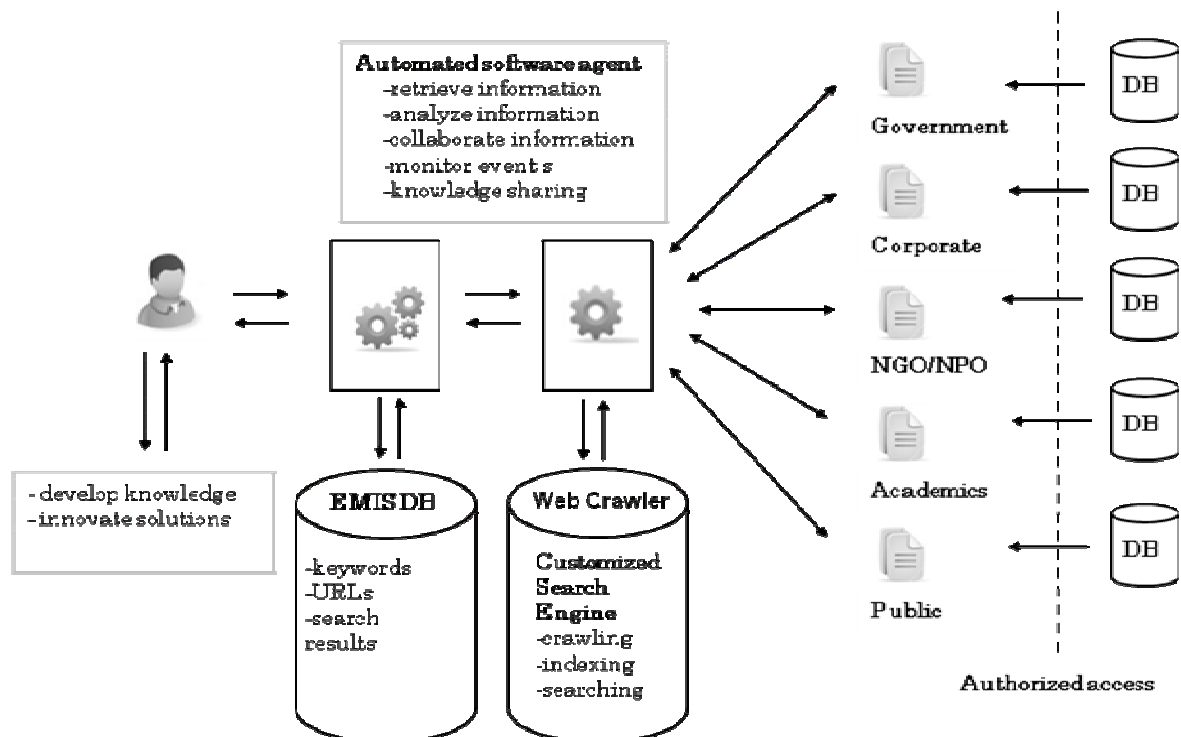


Figure 4: Information retrieval using EMIS with data crawling techniques

Regarding to the analysis and data mining function of the intelligent software agent, we must realize about the complex nature of structured and unstructured data of the Web and we need to integrate our proposed method with a proper data extraction or data mining software in order to achieve the information we are looking for.

6. Prototype Crawling Method

A simple prototype model has been created using Google Spreadsheets and Google Custom Search Engine (CSE) (Ref. Google CSE) to demonstrate the proposed concept of using intelligent software agent for automatic information retrieval and information collaboration of EMIS.

A merit of using Google Spreadsheet is its widespread use free of charge. Especially, when it comes to collecting data from small companies and from less developed countries, Google framework provides a practical solution to data crawling, as we explain in here.

In this scenario, it is assumed that Google Custom Search Engine as a focused Web crawler and Google Spreadsheets as automated software agent.

First, a custom search engine from Google is built using the seed URLs and preferred keywords related to interested environmental issues such as environment management", "global warming", "climate change", "greenhouse gases", "CO₂ emission". (See Figure 5). The Google Custom Search Engine returns more focus results from a site or collection of sites that we would like to search over.

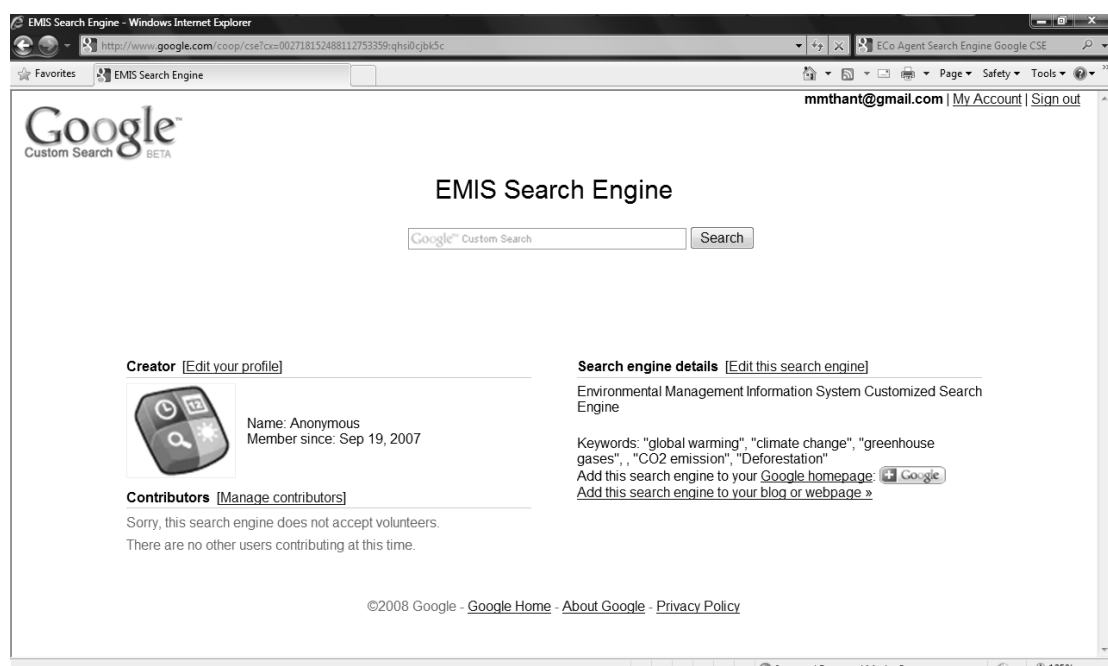


Figure 5: Google custom search engine

Second, we tried to extract and store the URLs of search results in a Google Spreadsheets. The results can be easily and automatically retrieved using the ImportXML(URL,query) function where URL is the searching URL of our Google Custom Search Engine with keyword as search string and query is the XPath syntax of searching hyperlinks //a[@class='l']/@href.

The complete syntax of the ImportXML function is as follows:

```
=ImportXML("http://www.google.com/cse?cx=
002702689857330478473%3A97ad2lazyqc&i
e=UTF-8&q=greenhose+gas+emission+per+c
```

Third, we analyze the URLs and tried to datamine the results which contain data in simple HTML table format. In the Google Spreadsheets, data table from the links can be extracted using ImportHTML (URL,element,Index) function. As an example we retrieved data table of green house gas emission from the Japan Center for Climate Change Action (JCCCA) website using import function (See Figure 6) and generate the graph (See Figure 7) (Ref. JCCCA GHG Data).

File Edit View Format Insert Tools Form Help										
<div> <div>\$ % 123 10pt B Abc A</div> <div> </div> <div>Σ</div> </div>										
	A	B	C	D	E	F	G	H	I	J
		二酸化炭素 (CO2)	メタン (CH4)	一酸化二 窒素(N2O)	ハイドロフ ルオロカー ボン (HFCs)	パーフル オロカー ボン (PFCs)	六フッ化 硫黄 (SF6)	合計	対基準 年増 減%	対前年増 減%
2	基準年	1144.1	33.4	32.6	20.2	14	16.9	1261.3		
3	1990	1144.2	32.6	32				1207.8	-4.2%	
4	1991	1152.6	32.4	31.5				1216.5	-3.6%	0.7%
5	1992	1160.8	32.1	31.5				1224.5	-2.9%	0.7%
6	1993	1153.6	31.8	31.3				1216.7	-3.5%	-0.6%
7	1994	1213.5	31.1	32.5				1277.1	1.2%	5%
8	1995	1226.6	30.2	32.8	20.3	14.1	17	1341.2	6.3%	5%
9	1996	1238.9	29.5	33.9	19.9	14.9	17.5	1354.7	7.4%	1%
10	1997	1234.9	28.5	34.6	19.9	16.3	15	1349.1	7%	-0.4%
11	1998	1198.9	27.6	33.1	19.4	13.5	13.6	1306.2	3.6%	-3.2%
12	1999	1233.9	27	26.7	19.9	10.6	9.3	1327.5	5.2%	1.6%
13	2000	1254.6	26.4	29.3	18.8	9.7	7.3	1346	6.7%	1.4%
14	2001	1238.8	25.6	25.8	16.2	8.1	6	1320.5	4.7%	-1.9%
15	2002	1276.7	24.7	25.5	13.7	7.5	5.7	1353.7	7.3%	2.5%
16	2003	1283.9	24.2	25.2	13.8	7.3	5.4	1359.7	7.8%	0.4%
17	2004	1282.5	23.8	25.3	10.6	7.5	5.3	1355	7.4%	-0.3%
18	2005	1287.3	23.4	24.8	10.6	7.1	4.6	1357.8	7.7%	0.2%
19	2006	1270.2	23	24.7	11.6	7.4	5.1	1342.1	6.4%	-1.2%
20	2007	1303.8	22.6	23.8	13.2	6.5	4.4	1374.3	9%	2.4%

Figure 6: Data scraping from Japan Center for Climate Change Action (JCCCA)

```
=importHTML("http://www.jccca.org/content/view/1043/784/", "table", 3)
```

The results can also be published the web page and can be shared globally.

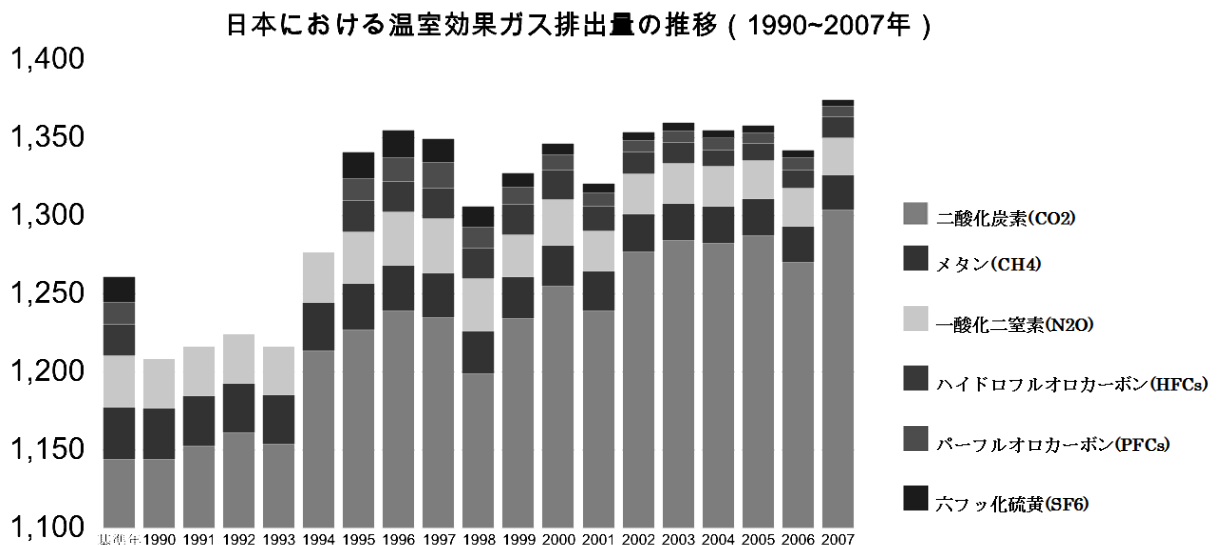


Figure 7: Display the data scraping results from Japan Center for Climate Change Action (JCCCA) in Google Spreadsheet chart

If the data files which are shared in .csv format, these files can be easily imported directly into the Google Spreadsheets using the ImportData(URL) function.

For example:

```
ImportData("http://cdiac.esd.ornl.gov/ftp/ndp030/CSV-FILES/global.1751_2005.csv")
```

In addition, the unique features of Google Spreadsheets such as real-time editing and collaboration can be used to share data among people without the need to send or download a file. Moreover, the data from Google Spreadsheets can be published to the Web or can be mashed together on a map.

7. Conclusion

From the results of our prototype development, we would like to discuss the practicability of using data crawling techniques in EMIS. We assume that it is feasible to develop an environmental management information system with data crawling and

intelligent data extraction software agents; although there are some constraints in developing intelligent focused crawler and information extractor.

Nevertheless, it is clear that information about global warming and the public awareness play a critical role and we need an effective data crawling and collaboration tool that enable the users to share the environmental data and information. As we have discussed in our prototype model, Google Spreadsheets could be one of the solutions; especially usable in developing countries and also by small and medium size companies even in developed countries. Once data is entered using Google Spreadsheets, following agreed upon data tags, a crawler as we described in this prototype, can be easily used to collect data and put into a database for further analysis, compiling, and reporting.

To sum up, if we have effective EMIS, we will understand more about the current environmental issues and we can develop innovative solutions to address these problems.

References

1. CAIT, About CAIT - Climate Analysis Indicators Tools; Website (May, 2009):
<http://cait.wri.org/faq-about-cait.php#5>
2. David Pinder and Brian Slack, "Shipping and ports in the twenty-first century: globalisation, technological change and the environment", Routledge, 2004.
3. Globalis, "Greenhouse Gas Emissions per Capita"; Website (May 2009):
<http://globalis.gvu.unu.edu/?2275>
4. Google CSE, Google Custom Search; Website (May 2009):
<http://www.google.com/coop/cse/>
5. IGES, "IGES EnviroScope-Online Platform on Environmental Strategy, Policy and Research"; Website (May 2009):<http://www.iges.or.jp/en/database/index.html>
6. JCCCA GHG Data, Import Data from JCCCA website to Google Spreadsheet; Website (May 2009):
<http://spreadsheets.google.com/pub?key=rAiQ5YWIq1ruDBty5iVSnoQ&output=html>
7. Jurgen Kropp and Jurgen Scheffran "Advanced methods for decision making and risk management in sustainability science", Nova Publishers, 2007.
8. NEC Corp., Japan, "Information Systems that Support Environmental Management"; Website (May 2009): <http://www.nec.co.jp/eco/en/03/3-15-01.html>
9. ODP, Open Directory Project; Website (May, 2009): <http://www.dmoz.org/>
10. Toshiba Corp., Japan, "Environmental Management Information System"; Website (May 2009):
http://www.toshiba.co.jp/env/en/management/information_system.htm
11. UNFCCC1: National Reports, United Nations Framework Convention on Climate Change website information; Website (May, 2009):
http://unfccc.int/national_reports/items/1408.php
12. UNFCCC2: Greenhouse Gas Inventory Data, United Nations Framework Convention on Climate Change website information ; Website (May 2009):
http://unfccc.int/ghg_data/items/3800.php
13. UNFCCC3: Kyoto Protocol, United Nations Framework Convention on Climate Change website information ; Website (May 2009):
http://unfccc.int/kyoto_protocol/items/2830.php
14. US EPA, "Environmental Management Systems (EMS)"; Website (May 2009):
<http://www.epa.gov/EMS/>
15. Web Crawler; Website (May 2009):
<http://www.webcrawler.com/info.wbcrawl.toolbar/search/help/about.htm>