OAI-PMH Repository Enhancement for the NASA Langley Research Center Atmospheric Sciences Data Center

Martin Klein
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
mklein@cs.odu.edu

Michael L. Nelson
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
mln@cs.odu.edu

Juliet Z. Pao
NASA Langley Research Center
Hampton, VA 23681
juliet.z.pao@nasa.gov

ABSTRACT

The NASA Langley Research Center Atmospheric Science Data Center (ASDC) has almost 2 petabytes of canonical earth science data. Despite this volume of data, the ASDC does not maintain the information resources derived from this data: publications, web pages, visualizations, etc. We present the preliminary results of a project that augments the holdings of the ASDC by using search engine APIs to find relevant resources on the web, determine their importance by computing their weighted position in various search engines, download the resources and package them in an XML format (MPEG-21 Digital Item Declaration) suitable for dissemination via an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) repository. We present the current status of this project and our plans for future work.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: [Digital Libraries]

General Terms

Design, Management

Keywords

Digital Libraries, Search Engine API

1. INTRODUCTION

The NASA Atmospheric Science Data Center (ASDC)[1] maintains an Earth Science Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH) [9] repository that contains earth science data consisting of 42 science

---

[1]http://eosweb.larc.nasa.gov/

projects with over 1700 data sets and 2M data granules in a combination of almost 2 petabytes of online and nearline storage. While the ASDC has the canonical technical and administrative metadata for the projects, collections and granules, the ASDC has very little in the way of descriptive metadata. Most of the descriptive metadata that does exist exists only in semi-structured or unstructured HTML pages and not in the ASDC repository itself. Consequently, the ASDC has an interest in discovering web resources that describe the earth science projects for which they curate data. From these web resources, they can generate a textual corpus that describes the project: metadata, word frequency, etc. They can also save timestamped local copies of the web resources so a full context of relevant information can be preserved for future access or analysis.

In this project we focused on discovering additional data and metadata pertaining to ASDC projects by utilizing search engines like Google, Yahoo! and MSN. The project also focuses on providing the discovered results in an archive ready format that is compatible with the NASA ASDC OAI-PMH repository.

2. RELATED WORK

This work is based on previous efforts by Chu et al. [3] to implement an OAI-PMH compliant repository at the NASA ASDC. Focused crawling [2] has traditionally been the primary method to gather online resources relevant to a specific topic. While many digital libraries still use this technique, focused crawling is now often augmented by using search engine APIs [8, 12, 4]. McCown et al. [7] use the search engine APIs to retrieve cached copies of resources to reconstruct lost web sites. [5]. They use SEs to find buying guides on the Internet. The authors find that those guides are often hard to find for the average user due to their rather simple (user generated queries). They build what they call "carnivores" that issues machine-generated queries to SEs and find buying guides for the user. Many researchers use the provided APIs, even though the results are often inconsistent with the web interface [6].

3. THE CONCEPT

Figure 1 displays the general idea behind the project. Instead of gathering pages with focused crawling, for each project we use the project's title and (if available) the acronym to issue a query to the three major Internet search engines
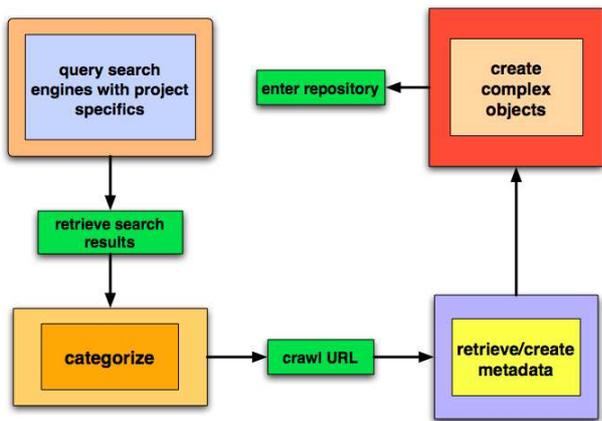
Figure 1: The General Concept of the Project

(SEs). We thereby take advantage of the power and ubiquity of SEs plus distribute the load of gathering information over powerful servers, away from a small scale focused crawler that we alternatively had to maintain.

Sending queries to the APIs of the three SEs returns three result sets that need to be merged. If we just consider the top ten results from each SE we have a total of 30 URLs. We do not know exactly how the SE rank their results and so we can only create an aggregated result set based in relative ranking. Therefore we calculate the weight for each URL (W_{url}):

$$W_{url} = \sum_{s \in SE} (B - R(s)) + 1 \qquad (1)$$

where B is the amount of top results we are using (e.g.10) and $R(s)$ is the rank of the returned URL in the result set of the SE it was retrieved from (s). This results in a complete list of (unique) retrieved results, ranked by their relative weight. Next we perform a very low level categorization by distinguishing between *PDF*- and *HTML*-documents by analyzing the URLs and the actual path. This is important for the next step where we crawl all HTML documents and store them on a local drive. We provide two sets of metadata:

1. existing metadata which we grab from the headers of the downloaded HTML files like the title of the page and all included meta tags

2. additionally created metadata like the timestamp of the crawl, the URL where the resource was crawled from and of course the computed weights of the current result set.

The downloaded resource is compressed in the file system and together with both sets of metadata eventually included in a complex object. We use MPEG-21 Digital Item Declaration Language (DIDL) [1] to encode the object. For redundancy and preservation reasons we include the data object itself both, by value and by reference into the DIDL object.

4. THE ARCHIVABLE OBJECT

As mentioned in section 3 we produce archive-ready complex objects which contain metadata about the discovered

resources as well as the crawled websites. The structure of the MPEG-21 DIDL complex objects and the representation of the records in the ASDC OAI-PMH repository will be covered in the next two sections.

4.1 The Complex Object

Figure 2 shows the structure of the MPEG-21 DIDL complex object. We are creating a new separate container for each ASDC project. The container project's acronym (PA) as the unique identifier. is identified by the project's acronym (PA) which in included in a descriptor statement of the container. Another descriptor in the container holds a timestamp which is set at the time of the creation of container. A DIDL container hosts all *items* for a particular ASDC project. An item represents one URL retrieved from querying the search engines meaning a container holds the same amount of items as URLs were returned. The PA plus the actual URL serves as the unique identifier for each item and a descriptor option with the timestamp of the item's creation time is added to each item as well. All items contains two more items each, one for the metadata set pertaining to the URL and one for the data objects. The metadata item is identified by the PA plus the appropriate URL plus the string "METADATA". The identifier of the data item is similar, it consists of the PA plus the according URL plus the string "DATA". Both items also hold the descriptor option with the creation timestamp. The metadata item contains a *component* which holds the metadata in a *resource*. The data item in contrast contains a component with two resources, one holds the object by reference and the other one holds the actual object by value. By definition, resources in an item are byte-wise equivalent.

We would like to stress that the project's acronym is not sufficient as an unique identifier for the long term. However, it was sufficient for our small scale prototype implementation and due to the flexibility of the system this setting can be changed at any time and thus meet large scale demands.

4.2 Representation in the Repository

As mentioned earlier the ASDC maintains an OAI-PMH compliant data repository where the MPEG-21 DIDL objects will be ingested. The representation the object in the repository is different from the structure of the complex object and thus will be addressed here.

Conceptually the program is intended to be rerun on a regular basis which involves issuing the same queries to the search engines again in order to discover new relevant data and possibly update the existing records. For preservation purposes the repository is supposed to be able to keep the originally crawled objects and also provide a history of updates made to the objects. At the same time it needs to avoid growing too big by adding objects with just minor updates or changes of the content compared to an earlier version of the object. For the purpose of usability it would be a great asset if the repository could provide a rollback function that allows the user to access objects that have been ingested in the past.

Figure 3 shows the representation of three objects of the "International Satellite Cloud Climatology Project" *ISCCP*[2] in the repository. At the time of crawling, meaning the querying of search engines, the timestamp that is added to the project's acronym (ISCCP) for the identifier is created.
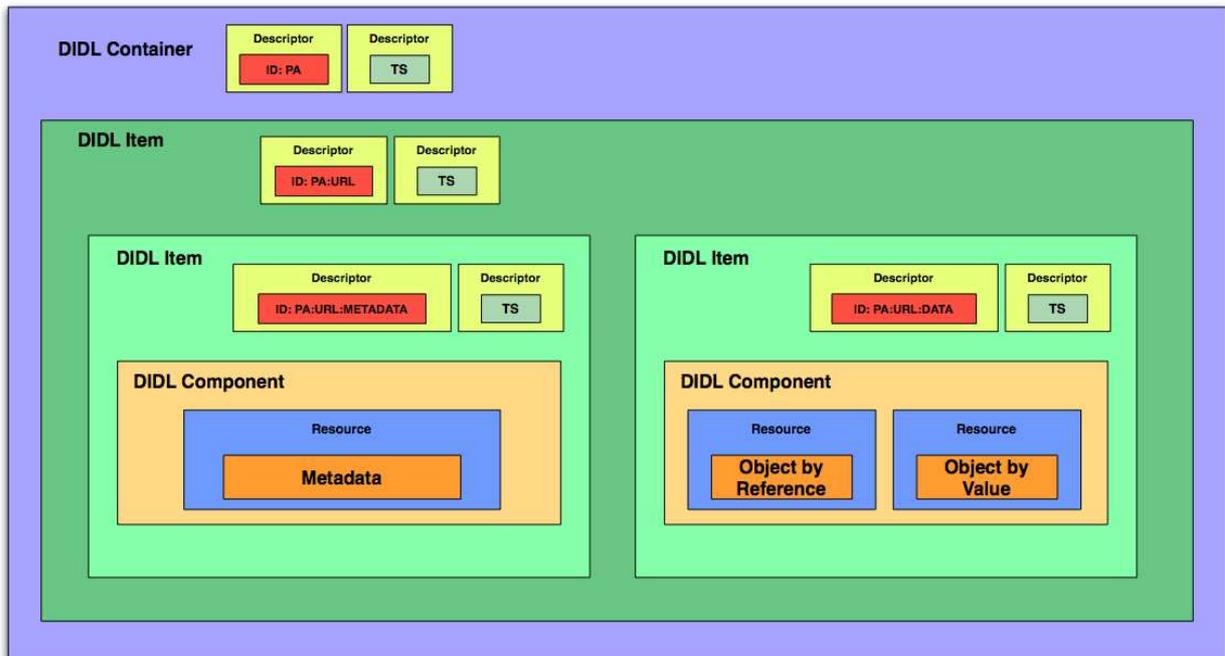
---

[2]http://isccp.giss.nasa.gov/

Figure 2: The Structure of the MPEG-21 DIDL Complex Object

The first crawl happens at time *TS1* and thus the identifier of the item is *ISCCP:TS1*. This ID does not change over time for the very same object. The *set* for all records is the same, determined by the project's acronym. Using the OAI-PMH filter option *set* enables the user to retrieve all objects related to a particular project at once. A list of all available projects could be retrieved by using *ListSets* verb.

Another timestamp (t_a) is set for the record which represents the time the record was created or modified. That means in case of modifications of the content which could include minor fixes, updates or even transitions in format, the ID does not change but the record timestamp does. This is shown in Figure 3 as *Modification1*, *Modification2* and *Modification3* and the changing timestamps of the records t_b, t_c and t_d. The optional OAI-PMH arguments *from* and *until* allow selective harvesting based on timestamps.

A rerun of the program can happen any time or following a schedule. In either way a new item in the repository is created with a new unique identifier containing the new timestamp. The set remains the same, depending on the project and the record's timestamp is reset to the time of the creation. This scenario is displayed in Figure 3 as *Crawl2 Modification1* and *Modification2*. Note the change of the ID *ISCCP:TS2* and record timestamps t_α, t_β and t_γ. Similar behavior can be seen for *Crawl3*.

The user seeking the most recent item of a particular project needs to issue a *ListIdentifiers* verb with the argument *set* and the project's acronym as the parameter. The timestamps in the returned IDs need to be compared to determine the most recent item. This follows the OAI-PMH philosophy of taking the burden of implementing complex functions that may increase usability away from the repository and putting them onto the harvester's side.

It also can be seen in Figure 3 that the harvester is able to access resources with older timestamps but are still maintained and available in the repository. A crawl can happen any time and so can a modification which means an item crawled and first ingested can be updated long after new items for the same project have been created. Thus the repository grows in breadth only when new items are created and not for every single minor update in the content.

As discussed in Chu et al. [3], the ASDC repository holds metadata records for (in decreasing generality) projects, collections and granules. Projects and collections are available in Dublin Core as well as Directory Interchange Format (DIF), while the granules are available only in DIF. The complex objects we create will be added as new metadata format (proposed as "asdc_crawl_archive"), with the OAI-PMH identifiers constructed as described above. Dublin Core metadata records for these identifiers will consist of the metadata extracted from the resources acquired in the results.

5. THE SYSTEM IMPLEMENTATION

We have implemented a prototype which can be run as a cron job on UNIX based systems. This ensures that the program can run following a schedule and keep the data in the repository up to date. The program is entirely written in *Perl* and uses CPAN modules that provide functions to parse HTML documents, write XML and query the SE's API.

As mentioned in section 3, the program creates a query string which consists of all terms of the project title plus its acronym if available. The terms are not enclosed by quotation marks which leaves it up to the SE to ignore terms that are too general like "the" and "a" for the search. The query string further does not contain a logical AND so that the results can contain any (compared to all) of the query
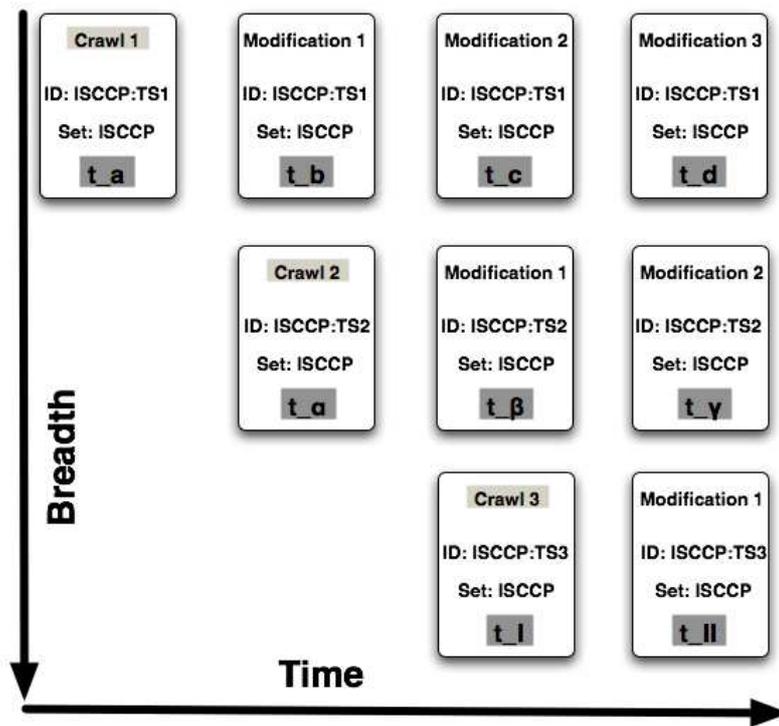
Figure 3: OAI-PMH Model

terms. It then automatically sends this query to the three major search engines (Google, Yahoo and MSN) utilizing their APIs. Yahoo!'s API can be used with the appropriate Perl module and Google and MSN provide a SOAP based API which makes it very convenient to issue queries and retrieve the results. The number of results returned from the APIs is configurable, in our initial experiments we chose to use the top ten and top 25 results from each SE.

All retrieved URLs are weighted using equation 1. For example, if the program queries the top ten search results from each SE the top ranked URL from each result set would get 10 points, so the maximum possible weight in this case would be 30. A URL that was ranked second in the Google result set, third in the Yahoo result set and fifth in the MSN result set would get scores of 9, 8 and 6 and end up with an overall weight of 23. This list of aggregated weighted URLs also serves the purpose of deleting duplicates (because of the relative weight, duplicate URLs are deleted) plus a HTML file containing that list is automatically generated and copied to a web server directory, so that a user can always access the file and browse through the result set while the actual program is running.

The prototype further separates HTML from non HTML documents and writes the results into two files one for HTML documents and one for non HTML documents. Both files are also copied to the web server directory. This enables the user to keep track of what URLs are going to be crawled in real time. The depth of the actual crawl is freely configurable. It was sufficient for our project crawler to follow just one level of links as long as they point to pages in the same domain.

The crawler uses the *Wget* program to download the doc-

uments and creates a new directory for each successfully crawled URL. While crawling, the prototype keeps track of the crawled URL and the current local time. This data is used as additional metadata records for the final repository object. After downloading the content, the program parses the files and extracts existing metadata, such as the title of the HTML documents and everything that can be found within HTML META tags. In order to save space each created directory is tar'ed and compressed. To enable the user to manually check the downloaded content, each directory (uncompressed) is also copied to the web server directory.

As the last step the prototype creates the complex object in the MPEG-21 DIDL format. The metadata sets gained from the downloaded content plus the additionally created data is embedded in the complex object. We also include the originally created list of weighted results from the SE as a metadata set which may be of special interest for users in the future. The data object is included in two different ways. The compressed directory is *base64* encoded and included into the complex object (by value) and the path to the web server directory which contains the crawled data is also included (by reference).

The prototype does not handle the process of ingesting the objects into the NASA ASDC repository.

6. RESULTS

The results of our experiments for example for the International Satellite Cloud Climatology Project (ISCCP) project are promising. We retrieved 23 unique URLs (the top 5 results appeared in the result set of more than one SE) while asking for the top ten results only. Using the top 25 results for the same project, we retrieved 64 unique URLs (overlap

of 11) with the maximum weight of 50. Out of this result set we can extract 39 unique top level domains (16 .gov, 8 .edu, 5 .org, 3 .com and 7 cc) which implies a great variety in the results. The table in the appendix shows the list of the 64 retrieved results with their computed weights and which SE they were returned from. Our prototype just crawls the first level of the retrieved URLs, meaning it does not follow any links into the web site's hierarchy but it downloads enough data to display the main page of the resource offline. Despite this shallowness, downloading the 39 unique URLs already adds up to a size of 8.7MB in the file system (4.2MB compressed).

Table 1 shows parts of the metadata we derived from parsing through the crawled content of the top ranked URL from our weighted list (http://isccp.giss.nasa.gov/). It becomes visible that the HTML sources indeed contain metadata which is worth ingesting into the complex object. We observed that some websites even follow the Dublin Core [11, 10] standard and label meta tags. In this example we find rather descriptive metadata like the title, description and keywords but also very specific technical and administrative metadata like the web master and the org code which may even be helpful for further internal identification. The values for the identifier, creation date and resource by reference were added by our system.

7. FUTURE WORK

Perhaps more important than discovering general resources on the web would be querying digital libraries to discover ASDC project related publications. We know relevant publications exist from performing manual searches and so we will utilize resources like the Smithsonian/NASA Astrophysics Data System[3], ScienceDirect[4] and of course also resources like Google Scholar[5], Citeseer[6] and Web of Science[7].

We will also take a closer look at the non-html results we get from the search engines. That includes parsing *PDF*-documents and retrieve relevant references (DOIs) and discover further related publications by crawling the author's and co-author's websites. A related problem is the categorization of results from search engines. Analyzing the extension of the filename in the URL is insufficient because MIME types do not always map to semantic types (e.g., not every PDF is an eprint).

Another open question is the problem of precision and recall. We observe a high precision in the top ten results given the title and acronym of a project. As a next step we will develop an automated or semi-automated way to evaluate how deep in the result list we can go to increase recall while making sure high precision is maintained. The question is, how far can we go in the result set before the results become irrelevant — this is likely to be evaluated by the ASDC administrators. We could take a hand-selected set of relevant resources as picked by ASDC administrators to establish "aboutness" for a project and reject results that fall below an empirically determined similarity threshold. Recall is difficult to solve because we are unsure of how much relevant material is available on the Internet.

---

[3]http://ads.harvard.edu/

[4]http://www.sciencedirect.com/

[5]http://scholar.google.com/

[6]http://citeseer.ist.psu.edu/

[7]http://www.isiwebofknowledge.com/

The process of automatically ingesting the created complex objects into the ASDC repository will be part of the future work. The program is meant to run automatically as a UNIX cron job.

8. CONCLUSION

We have developed a concept to automatically retrieve additional relevant data pertaining to a given ASDC project. We utilize three major Search Engines and aggregate their results. We also developed a prototype that is able to categorize the data and rank the results according to a relative weight that is computed for each of them. The returned URLs are crawled and stored on a local system. Existing metadata items are distilled by parsing the downloaded content and additional metadata such as the list of computed relative weights is created by the prototype. All metadata sets plus the data sets from a MPEG-21 DIDL complex object where the data is included by value (compressed files) and by reference (link to a location accessible via the Internet). We developed a creation and update policy including the structure for the complex object so that this archive ready complex object can eventually be ingested into the NASA ASDC OAI-PMH repository.

9. ACKNOWLEDGMENTS

10. REFERENCES

[1] J. Bekaert, P. Hochstenbach, and H. Van de Sompel. Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine*, 9(11), November 2003. doi:10.1045/november2003-bekaert.

[2] D. Bergmark. Collection Synthesis. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 253–262, 2002.

[3] C. Chu, W. E. Baskin, J. Z. Pao, and M. L. Nelson. OAI-PMH Architecture for the NASA Langley Research Center Atmospheric Science Data Center. In *ECDL '06: Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries*, pages 524–527, 2006.

[4] T. G. Habing, T. W. Cole, and W. H. Mischo. Developing a Technical Registry of OAI Data Providers. In *ECDL*, pages 400–410, 2004.

[5] R. Kraft and R. Stata. Finding Buying Guides with a Web Carnivore. In *Proceedings of LA-WEB*, pages 84–92, 2003.

[6] F. McCown and M. L. Nelson. Agreeing to Disagree: Search Engines and their Public Interfaces. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2007.

[7] F. McCown, J. A. Smith, and M. L. Nelson. Lazy Preservation: Reconstructing Websites by Crawling the Crawlers. In *Proceedings of WIDM*, pages 67–74, 2006.

[8] G. Pant, K. Tsioutsiouliklis, J. Johnson, and C. L. Giles. Panorama: Extending Digital Libraries with Topical Crawlers. In *JCDL '04: Proceedings of the 4th*

Table 1: Aggregated Metadata for the ISCCP Project

identifier	ISCCP:isccp.giss.nasa.gov
creation date	Tue Feb 13 19:12:17 2007
title	INTERNATIONAL SATELLITE CLOUD CLIMATOLOGY PROJECT
keywords	CLOUDS SATELLITE CLOUD DATA SATELLITE CLOUD CLIMATOLOGY DATA CLOUDS CLIMATE
description	THE FOCUS OF THE INTERNATIONAL SATELLITE CLOUD CLIMATOLOGY PROJECT IS TO COLLECT WEATHER SATELLITE RADIANCE MEASUREMENTS AND TO ANALYZE THEM TO INFER THE GLOBAL DISTRIBUTION OF CLOUDS THEIR PROPERTIES AND THEIR DIURNAL SEASONAL AND INTER ANNUAL VARIATIONS.
web master	Robert.B.Schmuck.1
no	Larry.D.Travis.1
content-owner	Ely.N.Duenas.1
orgcode	611
resource by reference	http://foo.bar.edu/isccp.giss.nasa.gov.tar.gz

ACM/IEEE-CS Joint Conference on Digital Libraries, pages 142–150, 2004.

[9] H. Van de Sompel, M. L. Nelson, C. Lagoze, and S. Warner. Resource Harvesting within the OAI-PMH Framework. *D-Lib Magazine*, 10(12), 2004.

[10] S. Weibel. Metadata: The Foundations of Resource Description. *D-Lib Magazine*, 1(1), 1995.

[11] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin Core Metadata for Resource Discovery, Internet RFC-2413, September 1998.

[12] Z. Zhuang, R. Wagle, and C. L. Giles. What's There and What's not?: Focused Crawling for Missing Documents in Digital Libraries. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 301–310, 2005.

APPENDIX

#	url	weight	resource[rank]
1	http://isccp.giss.nasa.gov/	50	Google[1], Yahoo[1]
2	http://www.gewex.org/isccp.html	47	Google[3], Yahoo[2]
3	http://www.cgd.ucar.edu/cas/catalog/satellite/isccp/	40	Google[6], Yahoo[6]
4	http://www2.ncdc.noaa.gov/docs/isccp/icaweb.htm	29	Google[11], Yahoo[12]
5	http://www.ncdc.noaa.gov/oa/rsad/isccpb1/isccpproject.html	28	Google[9], Yahoo[15]
6	http://eosweb.larc.nasa.gov/GUIDE/campaign_documents/isccp_project.html	26	Google[19], Yahoo[7]
7	http://www.atmos.umd.edu/ srb/	25	MSN[1]
8	http://www.atmos.umd.edu/ srb/gcip/	24	MSN[2]
9	http://isccp.giss.nasa.gov/overview.html	24	Google[2]
10	http://eosweb.larc.nasa.gov/PRODOCS/srb/table_srb.html	23	MSN[3]
11	http://www.giss.nasa.gov/	23	Yahoo[3]
12	http://pubs.giss.nasa.gov/abstracts/1983/Schiffer_Rossow.html	23	Google[16], Yahoo[13]
13	http://www.gewex.org/srb.html	22	MSN[4]
14	http://www.gewex.org/ISCCP_data_products_4-2006.pdf	22	Google[4]
15	http://www.gewex.org/PAN-GEWEX-MTG/Pan-GEWEX_ISCCP-2006.pdf	22	Yahoo[4]
16	http://gcmd.nasa.gov/records/GCMD_LDEO_ISCCP.html	21	Yahoo[5]
17	http://www.cgd.ucar.edu/cas/catalog/satellite/isccp/C2/means.html	21	Google[5]
18	http://www.gewex.org/srbdata.htm	21	MSN[5]
19	http://islscp2.sesda.com/ISLSCP2_1/html_pages/groups/rad/srb_radiation_1deg.html	20	MSN[6]
20	http://eosweb.larc.nasa.gov/PRODOCS/isccp/table_isccp.html	20	Google[21], Yahoo[11]
21	http://islscp2.sesda.com/ISLSCP2_1/html_pages/groups/rad/srb_clouds_1deg.html	19	MSN[7]
22	http://citeseer.ist.psu.edu/18329.html	19	Google[7]
23	http://www.agu.org/eos_elec/95206e.html	18	MSN[8]
24	http://directory.eoportal.org/info_InternationalSatelliteCloudClimatologyProjectISCCP.html	18	Google[8]
25	http://dss.ucar.edu/datasets/ds742.0/docs/1983.SchifferRossow.pdf	18	Yahoo[8]
26	http://directory.eoportal.org/info_SurfaceRadiationBudgetSRBProject.html	17	MSN[9]
27	http://grp.giss.nasa.gov/links.html	17	Yahoo[21], MSN[14]
28	http://gcmd.gsfc.nasa.gov/records/GCMD_ISCCP_D1_NAT.html	17	Yahoo[9]
29	http://www.wmo.ch/web/sat/en/ap4-04.htm	16	Yahoo[10]
30	http://esrb.iesl.forth.gr/01project/01summary.htm	16	MSN[10]
31	http://daac.gsfc.nasa.gov/fieldexp/TOGA/isccp_dx.html	16	Google[10]
32	http://www.dwd.de/en/FundE/Klima/KLIS/int/CM-SAF/products/index.htm	15	MSN[11]
33	http://isccp.giss.nasa.gov/projects/flux.html	14	MSN[12]
34	http://www2.ncdc.noaa.gov/docs/isccp/cdcweb.htm	14	Google[12]
35	http://www.agu.org/pubs/crossref/1997/96JD03865.shtml	13	Google[13]
36	http://isccp.giss.nasa.gov/projects/gewex.html	13	MSN[13]
37	http://pubs.giss.nasa.gov/abstracts/2004/Rossow_Duenas.html	13	Google[23], Yahoo[16]
38	http://ghrc.msfc.nasa.gov:5721/campaign_documents/fire_ace.html	12	Yahoo[14]
39	http://www.agu.org/pubs/crossref/1996/96JD01771.shtml	12	Google[14]
40	http://microwave.nsstc.nasa.gov:5721/campaign_documents/fire_ace.html	11	Google[15]
41	http://grp.giss.nasa.gov/projects.html	11	MSN[15]
42	http://rredc.nrel.gov/otherlinks.html	10	MSN[16]
43	http://www.ghcc.msfc.nasa.gov/ampr/fire3.html	9	Yahoo[17]
44	http://www.iode.org/oceanportal/detail.php?id=5681	9	Google[17]
45	http://ams.confex.com/ams/Madison2006/techprogram/session_19727.htm	9	MSN[17]
46	http://asd-www.larc.nasa.gov/ceres/brochure/land_cover.html	8	MSN[18]
47	http://ams.allenpress.com/perlserv/?request=get-abstract&doi=10.1175%2FBAMS-85-2-167	8	Google[18]
48	http://www.cgd.ucar.edu/cas/catalog/satellite/isccp/D2/references.html	8	Yahoo[18]
49	http://badc.nerc.ac.uk/data/isccp_top/	7	Yahoo[19]
50	http://asd-www.larc.nasa.gov/people_html/stackhouse.html	7	MSN[19]
51	http://www.science.gmu.edu/ jwang/albedo/node1.html	6	MSN[20]
52	http://adsabs.harvard.edu/abs/1990rete.conf...60R	6	Google[20]
53	http://eobglossary.gsfc.nasa.gov/Library/GlobalClouds/cloudiness3.html	6	Yahoo[20]
54	http://www.dar.csiro.au/sensing/index.html	5	MSN[21]
55	http://grp.giss.nasa.gov/clouds_ref.html	4	Yahoo[22]
56	http://adsabs.harvard.edu/abs/2004BAMS...85..167R	4	Google[22]
57	http://ccrp.tor.ec.gc.ca/CAGES/assess/gewex_6.htm	4	MSN[22]
58	http://directory.eoportal.org/info_IDSDataSetsfortheInternationalSatelliteCloudClimatologyProject.html	3	Google[25], Yahoo[24]
59	http://www.sparc.sunysb.edu/html/data_links.html	3	MSN[23]
60	http://directory.eoportal.org/info_FIREArcticCloudExperimentACE.html	3	Yahoo[23]
61	http://www.clivar.org/organization/southern/other.htm	2	MSN[24]
62	http://dss.ucar.edu/catalogs/ranges/range712.html	2	Google[24]
63	http://www-cger.nies.go.jp/cger-e/db/info-e/InfoDBWeb/frames/prog2/gewex.htm	1	MSN[25]
64	http://www.cira.colostate.edu/climate/ISCCP/ISCCPSPC.HTM	1	Yahoo[25]