Where in the World? Demographic Patterns in Access Data

Mimi M. Recker[1], Beijie Xu[1], Sherry Hsi[2], and Christine Garrard[3]
mimi.recker@usu.edu
beijie.xu@aggiemail.usu.edu
sherryh@berkeley.edu
chrisg@leupold.gis.usu.edu
[1]Department of Instructional Technology & Learning Sciences, Utah State University
[2]Center for Technology Innovation, Lawrence Hall of Science, UC Berkeley
[3]RS/GIS Laboratory, Utah State University

Abstract. As learning environments become increasingly available online, the fine-grained records of user activities can be captured and analyzed (generally called *webmetrics)* to better understand the characteristics of teachers and learners. The purpose of this paper is three-fold. First, we describe how two web-based educational systems were engineered to collect webmetrics. This is followed by a description of our methods for collecting geo-referenced data, joining these data with demographic and educational datasets for the United States, and mapping these using geographic information system (GIS) techniques to visually display relationships. We conclude with results from statistical analyses of these relationships that highlight areas of significance.

1 Introduction

Web-based learning environments can be engineered to collect detailed data about their users and their activities. These data can be subsequently mined and analyzed against a set of criteria (generally called *webmetrics* [9, 20]) to find behavioral patterns of individual teachers and students and to infer their characteristics and teaching/learning profiles [2, 11, 13, 18]. Such inferences help evaluate and improve system design, as well as measure the effectiveness of e-learning and teaching [1, 4, 5, 14].

In addition, standard webmetric tools record the IP address of users' computers, thereby providing fodder for analyses of the geographical location of users. These geographically referenced access patterns can be compared to national datasets to see if any relationships exist with demographic and educational trends. In this way, comparisons can be made between *samples* of users as reflected in access patterns to the *population* of users.

The purpose of this paper is three-fold. First, we describe how two web-based educational systems were engineered to collect webmetrics. This is followed by a description of our methods for collecting geo-referenced data, joining these data with demographic and educational datasets for the United States, and mapping these using geographic information system (GIS) techniques to visually display relationships. We conclude with results from statistical analyses of these relationships that highlight areas of significance.

2 System Descriptions

The two web-based learning environments that were engineered to collect user data and user activities are the Instructional Architect, a digital library service, and the Exploratorium Learning Resources Collection, a digital library for K12 teachers.

2.1 The Instructional Architect

The Instructional Architect (IA.usu.edu) is an educational digital library service developed to support the instructional use of the National Science Digital Library (NSDL.org) [10] and other online learning resources. With the IA, teachers are able to search, select, sequence, annotate and reuse online learning resources to create instructional web pages, called IA projects. The IA breaks down the technology barriers and allows teachers with basic computer skills to create online projects and efficiently address their instructional needs [15, 16, 17]

The Instructional Architect targets two main audiences: teachers and their students. A teacher can create a free account, which provides exclusive access to his/her saved resources and IA projects. As part of the registration procedure, the teacher completes a profile indicating subjects and grades taught, teaching experience, and level of information literacy.

After a teacher logs into the system, the IA offers two major usage modes: 1) **resource management** and 2) **IA project management**. In the **resource management** mode, teachers can search for and store links to NSDL resources, as well as add non-NSDL resource links to their individual collection.

The **IA project management** mode allows teachers to create a web page (called an *IA project)* and share it with the public or only with their own students. JavaScript and HTML code is allowed, which means dynamic objects such as multimedia, blogs, and RSS can be included. A registered teacher can create a generic student account that is shared by all his/her students. With such an account, students have exclusive access to their teachers' private projects that are marked as "student-view only".

Table 1 shows summary statistics of IA usage and growth over the previous year.

Table 1. Usage data growth (to December 2009).

	Since	Number	Previous year growth (%)
Registered users	09/2004	4,700	36
Web resources used	01/2005	43,410	51
IA projects created	09/2003	9,653	53
Project views	08/2005	~1 million	67

2.2 The Exploratorium Learning Resources Collection

The Exploratorium Learning Resources Collection (ELRC; www.exploratorium.edu) is a digital library of over 700 teacher-tested science activities and instructional resources inspired and created from the Exploratorium[1]'s exhibits, public program events, and teacher professional development programs.

The ELRC is designed for elementary and secondary school teachers as its primary audience, and informal educators as its secondary audience. Teachers can browse the collection by topic or conduct keyword searches. Search results provide a short description of the item, as well as related topics to explore. For each resource item found, a resource record is provided that includes a description of the resource as well as teaching tips where appropriate. Advanced search enables teachers to narrow a search by curricular area, grade level, and specific resource type (i.e., image, video, activity, article, web interactive, web exhibition, museum exhibit, and professional development resource). The ELRC is available to use from the Web and no registration is required to use the application. Through interoperability, the ELRC items can be also found in other educational libraries including the National Science Digital Library (nsdl.org) [10].

In 2009, the ELRC was accessed by over 34,500 unique visitors. Since its launch in 2005, visitors from all 50 States across North America and over 170 countries have accessed the collection. The most frequently accessed resources include the website page from "Faultline" which explains the differences between P and S waves, and a hands-on electric circuit activity called "Jitterbug."

3 Methods

This section describes the three-part process for 1) collecting, 2) joining and mapping, and 3) analyzing location data.

3.1 Collecting Webmetrics

Three general approaches are used in industry for collecting and analyzing webmetrics data. First, server log parsers are software tools that analyze the traffic data from Web server logs. For example, the Exploratorium uses Summary (summary.net) as a web traffic data analysis tool [7, 19].

Second, web sites fueled by a database can be engineered with locally developed software to capture access traffic. For example, *ActiveMath*, an adaptive learning environment, has a database that stores not only raw data but also analyses of users' actions and additional background knowledge concerning the users [11].

The third approach, called page-tagging, uses third-party services that capture web traffic by embedding Javascript into web pages. Popular examples include Google Analytics (analytics.google.com) [3, 6] and Omniture (www.omniture.com).

---

[1] The Exploratorium is a hands-on museum of science, art, and human perception located in California.

Since mid-2006, the Instructional Architect has been engineered to collect detailed online usage data, using both Google Analytics (GA) and locally developed tracking software (e.g., pageviews, session lengths, user paths) to measure online user behavior [9, 21]. The ELRC main website has used Google Analytics since 2007.

Google Analytics, as part of their standard reporting tool, estimates the visitors' location using the client computers' IP addresses, in a process called geo-location [8, 12]. Visitors' location data can then be interactively partitioned into geographical regions such as countries, U.S. states, and cities using geographic segmentation. The remainder of this paper discusses how to use the geographic segmentation data generated by GA to study more global web usage patterns.

3.2 Joining and Mapping Location Data

GA location data can be joined with other geographical datasets. In the present paper, we extracted U.S. demographic data from two sources: 1) commercial GIS software ESRI Data & Maps 9.3 (www.esri.com) and 2) the National Center for Education Statistics (NCES; nces.ed.gov). Both interactively provide detailed demographic and educational data about U.S. population that can be overlayed with GA location data.

The ESRI Data & Maps 9.3 software package was used to generate basic maps and city location data, which were then linked with the geographic segmentation from Google Analytics and educational data from NCES. Once these data were joined, they were imported into ArcMap 9.3, a main component of ESRI's GIS software, for visual display.

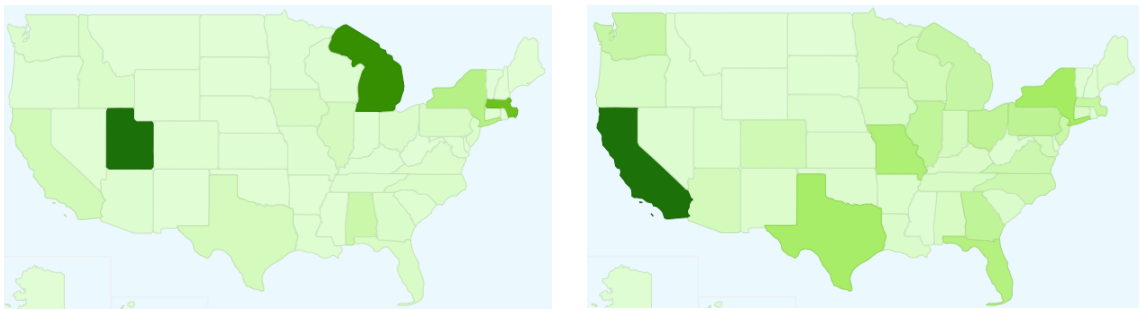
3.3 Analyzing Location Data

Finally, we examined statistical relationships between five demographic predictors and the number of site visits per location as reported by GA. The first predictor, 2007 U.S. population, was extracted from the ESRI dataset. The four additional U.S. demographic predictors, extracted from NCES, were: 2) number of schools, 3) number of school districts, 4) per capita income, and 5) median family income. Any number of predictors could have been selected, but given that computers are needed to access the IA and the ELRC from schools or homes, we focused on those related predictors.

4 Results

4.1 Collecting Location Data from GA

Using the GA reporting tool, Figure 1a&b shows visits to the IA and ELRC, respectively, over a 1-year period in the U.S. Darker shades indicate more visits. Given that the IA is based in Utah and has outreach activities in New York and Michigan, while the ELRC is California-based, the maps show that both groups are successful in local dissemination activities. The ELRC also shows more widespread U.S. visitors.



(a) The Instructional Architect (b) ELRC

Figure 1a&b. U.S. Geo-segmentation generated by Google Analytics. Darker shade indicates more visits.

4.2 Displaying GIS Mapping Data

Several maps were constructed to visually display the relationships between site accesses and demographic data. For example, Figures[2] 2a&b and 3a&b show IA and ELRC visitor traffic overlayed on two selected datasets (number of school districts and median family income). These maps help visually reveal relationships between site usage and demographic or school characteristics.

4.3 Analyzing Location Data

Of the 5 predictor variables described above, U.S. population is highly correlated with the number of schools (r = .97), and per capita income was highly correlated with median family income (r = .93). As such, two predictors, the number of schools and median family income, were removed from subsequent analyses.

Because certain states reported extremely high visits due to intense outreach or teacher professional development activities, data from such places were considered as outliers and were excluded to reduce measurement errors. Therefore, visits from Utah, Michigan, New York and Massachusetts were removed from the IA site, and visit data from California were removed from the ELRC site. Nonetheless, we observe unusual peaks in some localities, and such outliers were adjusted to the maximum. For both sites, values larger than 250 visits were adjusted to 250.

Even with this adjustment, the outcome variable, number of site visits per location, was extremely skewed, with the variance 50 times larger than the mean for the IA, and 100 times larger than the mean for the ELRC. Due to this over-dispersion, the Poisson regression does not fit for estimating the count distribution in our scenario. As such, we adopted the less restrictive model – negative binomial regression.

---

[2]For space reasons, only the continental 48 states are shown.

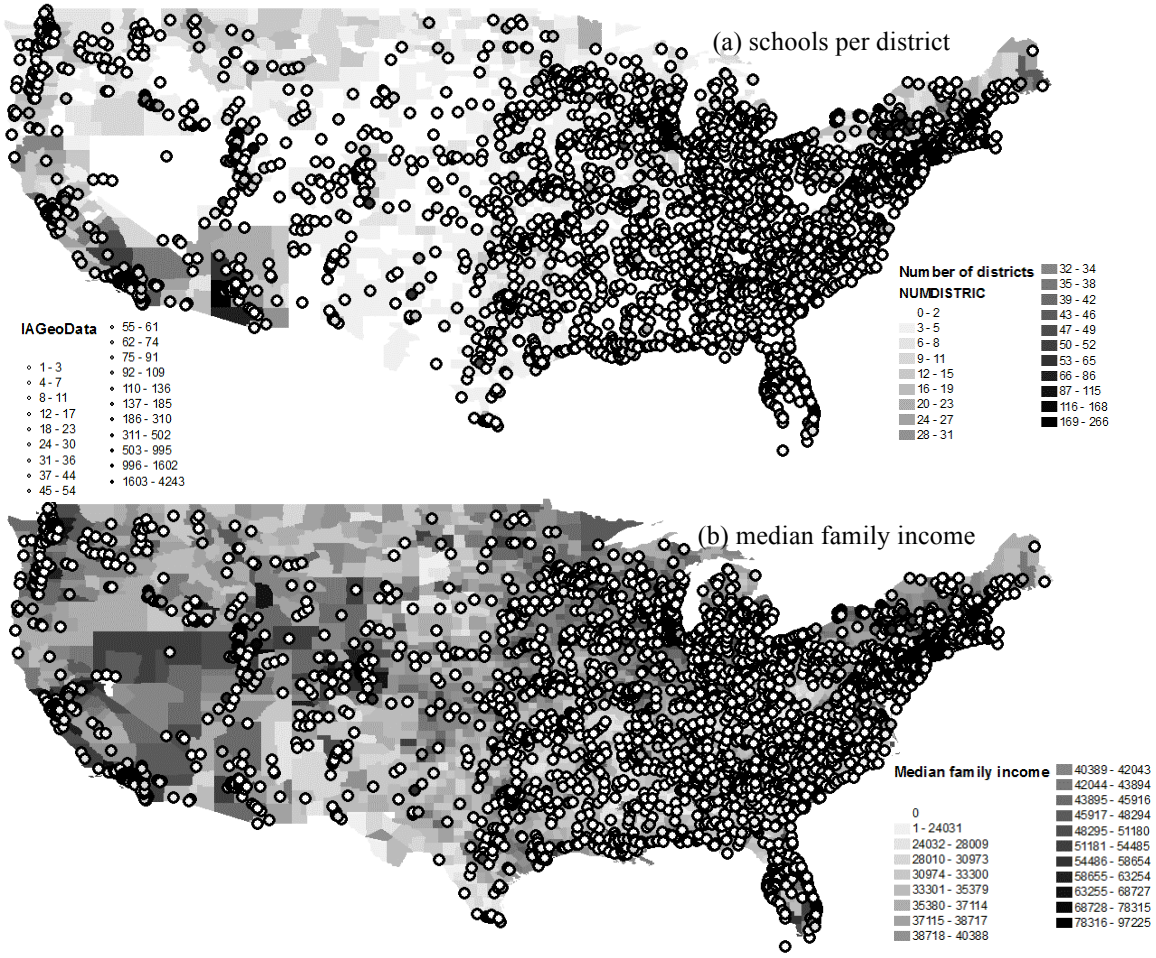


Figure 2a&b. U.S. map showing IA visits (darker dot is higher visit frequency) overlayed with number of school per districts, and (b) median family income over 1 year.

For IA, the negative binomial regression model predicting number of site visits per location from population and per capita income were each statistically significant, with *population* (Wald Chi-Square = 190.18, $p = .000$), and *per capita income* (Wald Chi-Square = 27.57, $p = .000$). The *number of school districts* was not (Wald Chi-Square = .63, $p = .43$).

For the ELRC, all three remaining indicators were statistically significant, with *population* (Wald Chi-Square = 71.36, $p = .000$), *per capita income* (Wald Chi-Square = 11.70, $p = .001$), and *number of school districts* (Wald Chi-Square = 6.96, $p = .008$).

We interpret these results to mean that online visitors to these sites came from, not surprisingly, more densely populated areas. In addition, the relationship with per capita income may be a function of the amount of resources (i.e., computing) available in the local schools and communities. Finally, online visitors to the museum's digital library

came from areas with a high number of school districts, while not such relationship was found with IA visitors.

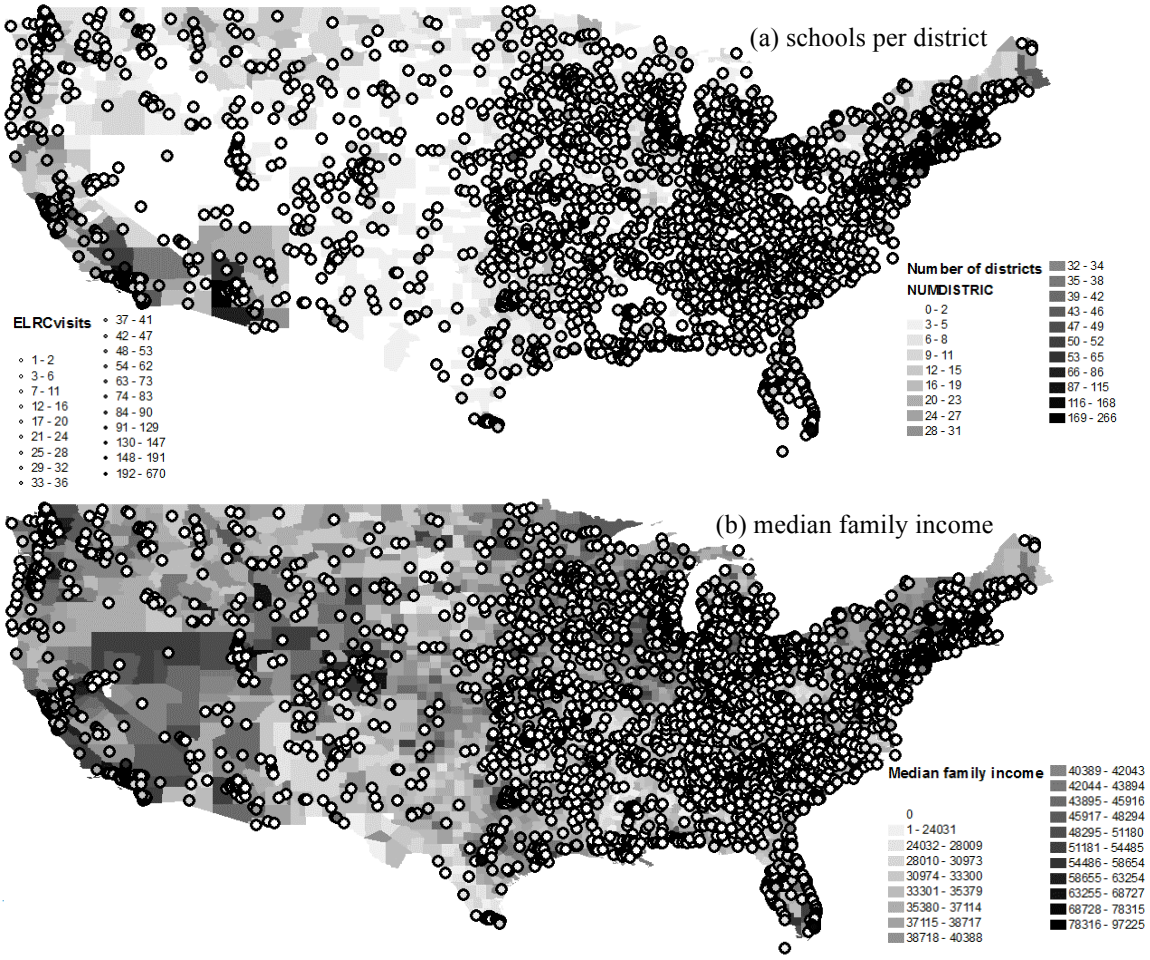


Figure 3a&b. U.S. map of ELRC visits (darker dot is higher visit frequency) overlayed with (a) number of schools per district, and (b) median family income over 1 years.

5 Conclusion

As learning environments become increasingly available online, the fine-grained records of user activities can be captured and analyzed to better understand user intent. Moreover, their online availability also provides access to users that can be beyond the initially targeted audience. GIS and statistical analyses of the geographical location of users collected from two online learning environments show interesting access patterns. First, visits are higher were outreach is higher, demonstrating the impact of these activities. Second, strong statistical relationships were found between certain U.S. demographic variables and location visit data. In particular, we did not necessarily expect to find the strong relationship between per capita income and the location of visitors. Is it possible that innovative online learning environments are exacerbating the digital divide?

In summary, this paper reports on both processes and results from one kind of analysis of webmetrics. Of course, these datasets contain a treasure trove of data mining potential for better understanding user activities. In current work, we are applying latent class analysis to usage data in order to help reveal different categories of users [21]. We also plan to triangulate and validate these findings with more conventionally collected data about user activities, including from user surveys and registration profiles.

6 Acknowledgment

This material is based upon work supported by the National Science Foundation under Grants No. 840745 & 0840738. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank Andrew Walker, Bart Palmer, Bob Donahue, Kerstin Schroder, and Yu-Chun Kuo.

References

[1] Borgman, C. L., Abelson, H., Dirks, L., Johnson, R., Koedinger, K. R., Linn, M. C., ... Szalay, A. *Fostering Learning in the Networked World – The Cyberlearning Opportunity and Challenge: A 21st Century Agenda for the National Science Foundation* (Report of the NSF Task Force on Cyberlearning), 2008. Arlington VA: NSF.

[2] Chan, S. Towards New Metrics of Success for On-line Museum Projects. In Trant. J. & Bearman, D. (Eds.), *Proceedings of the Museums and the Web 2008,* 2009. Toronto, Canada: Archives & Museum Informatics.

[3] Clifton, B. *Advanced Web Metrics with Google Analytics*, 2008. Indianapolis, Indiana. Wiley Publishing, Inc.

[4] Computing Research Association. *Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda,* 2005. Washington, DC. Available online at: www.cra.org/reports/cyberinfrastructure.pdf .

[5] Durfee, A., Schneberger, S., and Amoroso, D. L. Evaluating Students' Computer-based Learning Using a Visual Data Mining Approach. *Journal of Informatics Education Research*, 2007, 9(1), p. 1-28.

[6] Fang, W. Using Google Analytics for Improving Library Website Content and Design: A Case Study. *Library Philosophy and Practice 2007*: *LPP Special Issue on Libraries and Google*, 2007. Available online at: http://www.webpages.uidaho.edu/~mbolin/fang.htm.

[7] Hsi, S. Approaches to Collecting Online Audience Feedback and Measuring Impact. *Presentations on the Museum Computer Network Conference 2007*, 2007. Chicago, November 7-11.

[8] Katz-Bassett, E., John, J. P., Krishnamurthy, A., Wetherrall, D., Anderson, T., & Chawathe, Y. Toward IP Geolocation Using Delay and Topology Measurements.

Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, 2006. Rio de Janeriro, Brazil, pp. 71-84.

[9] Khoo, M., Recker, M., Pagano, J., Palmer, B., Washington, A., & Donahue, R. A. Using Webmectrics to Analyze Digital Libraries, *Proceedings of Joint Conference on Digital Libraries*, 2008. New York: ACM.

[10] McArthur, D.J. & Zia, L.L. From NSDL 1.0 to NSDL 2.0: Towards a Comprehensive Cyberinfrastructure for Teaching and Learning. *Proceedings of Joint Conference on Digital Libraries,* 2008, p. 66-69. New York: ACM.

[11] Muehlenbrock, M. Automatic action analysis in an Interactive Learning Environment. *Proceedings of the workshop on Usage Analysis in Learning Systems at AIED 2005*, 2005. Amsterdam, The Netherlands.

[12] Muir, J. A., & Oorschot, P. C. V. Internet Geolocation: Evasion and Counterevasion. *ACM Computing Surveys (CSUR),* 2009, 42(1). Article 4.

[13] Nickles III, G. M. Identifying Measures of Student Behavior from Interaction with a Course Management System. *Journal of Educational Technology Systems*, 2006, 4(1), p. 111-126.

[14] Pahl, C., & Donnellan, D. Data mining technology for the Evaluation of Web-Based Teaching and Learning Systems. *Proceedings of E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education*, Montreal, 2002, Quebec, Canada, 747-752.

[15] Recker, M. Perspectives on Teachers as Digital Library Users: Consumers, Contributors, and Designers. *D-Lib Magazine*, 2006, 12(9).

[16] Recker, M., Dorward, J., Dawson, D., Halioris, S., Liu, Y., Mao, X., ... Park, J. You Can Lead a Horse to Water: Teacher Development and Use of Digital Library Resources. In *Proceedings of the Joint Conference on Digital Libraries,* 2005, 1-9. NY, NY: ACM.

[17] Recker, M., Walker, A., Giersch, S., Mao, X., Halioris, S., Palmer, B., ... Robertshaw, M. B. A Study of Teachers' Use of Online Learning Resources to Design Classroom Activities. *New Review of Hypermedia and Multimedia, 2007*, 13(2), p. 117-134.

[18] Romero, C., & Ventura, S. Educational Data Mining: A Survey from 1995 to 2005. *Expert Systems with Applications,* 2007, 33(1), p. 135-146.

[19] Semper, R., Wanner, N., Jackson, R., & Bazley, M. Who's Out There? A Pilot User Study of Educational Web Resources by the Science Learning Network (SLN). *Paper presented at the Museums & the Web 2000 Conference*, 2000. Minneapolis, MN.

[20] Weischedel, B., & Huizingh, E. K. Website Optimization with Web Metrics: A Case Study. *Proceedings of the 8th International Conference on Electronic Commerce*, 2006, p. 463-470. Fredericton, New Brunswick, Cananda.

[21] Xu, B., Recker, M., & Hsi, S. The Data Deluge: Opportunities for Research in Educational Digital Libraries. In Evans, C. M. (Ed). *Internet Issues: Blogging, the Digital Divide and Digital Libraries,* 2010. Hauppauge, NY: Nova Science Publishers.