An Experimental Framework for Comparative Digital Library Evaluation: The Logging Scheme

Claus-Peter Klas
Norbert Fuhr
Sascha Kriewel
University of Duisburg-Essen

Hanne Albrechtsen
Institute of Knowledge Sharing

Laszlo Kovacs
Andras Micsik
SZTAKI

Preben Hansen
Swedish Institute of Computer Science

Giannis Tsakonas
Sarantos Kapidakis
Christos Papatheodorou
Ionian University

Elin Jacob
Indiana University
Bloomington

ABSTRACT

Evaluation of digital libraries assesses their effectiveness, quality and overall impact. To facilitate the comparison of different evaluations and for supporting the re-use of evaluation data, we are proposing a new logging schema that will account for all kinds of data about users, systems and the user-system interactions. We present a novel, multi-level logging framework that will provide complete coverage of the different aspects of DL usage. Based on this framework, we can analyse for various DL stakeholders the logging data according to their specific interests. In addition, specific analysis tools and a freely accessable log data repository will yield synergies and sustainablity in DL evaluation. The overall goal is to build a community for DL evaluation, based on agreement and discussion on a common ground.

Categories and Subject Descriptors: H.3.7 Information Storage and RetrievalDigital Libraries: standards, user issues

1. INTRODUCTION

Evaluation of digital libraries (DLs) aims at assessing their effectiveness, quality and overall impact. Analysis of transaction logs is one evaluation method that has provided DL stakeholders with substantial input for making managerial decisions and establishing priorities, as well as indicating the need for system enhancements. However, the quantitative nature of this method is often criticized for its inability to provide in-depth information about user interactions with the DL being evaluated. Because the results of a logging study are often localized and not easily interpretable outside the DL being investigated, recommendations arising from such an analysis are not easily generalizable to other DLs. The problem of generaliz-ability is compounded by the absence of a standardized logging scheme that could map across the various logging formats being used, thus presenting them in a commonly understandable language. The development of such a scheme would facilitate comparisons across DL evaluation activities and provide the means for highlighting critical events in user behaviour and system performance.

In our whitepaper on DL evaluation [1], we have identified the long-term goal of building a community for DL evaluation. Under the umbrella of an experimental framework that will serve as a joint theoretical and practical platform for the evaluation of DLs, the proposed logging scheme will allow for the meaningful interpretation and comparison of DL transactions.

In order to support the re-use of all possible evaluation data, we intend that the proposed scheme will account for all manner of data that can be collected from the user, the system and the user-system interaction. To this end, we are proposing a novel, multi-level logging framework that will provide complete coverage of the different aspects of DL usage. The main focus is the so called *concept level*, which generalizes events for comparison. Based on this specification, various DL stakeholders can analyze the logging data according to their specific interests.

A first proposal for a standard logging scheme for DLs was presented in [2]. However, when considering not only web based digital libraries, but a fully developed bibliographic application like DAFFODIL[1], with a rich set of services, the proposal has to be seen as a starting point. Therefore we push the issue further.

2. LEVELS OF LOGGING

When using transaction logs for evaluation, the main participants under survey are the user and the system, as well as the content that is being searched, read, manipulated, or created. The interaction between the system and the user can be examined and captured at various levels of abstraction, where each should have its own agreed-upon XML logging schema:

User behaviour level: Here the users and their behavior are located. Each user has a task to accomplish, within a certain social environment, and brings to that task her

[1] `http://www.daffodil.de`

JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.

individual knowledge.

Concept level for comparative evaluations: Here we locate generalized events generated by the DL user. By logging these events, user evaluation can be backed up with statistical data and a comparative evaluation of different users, systems and system content can be undertaken.

Service level: The service level represents a more specific set of system-dependent DL services. Because of differing input and output parameters, which are generalized on the concept level, service level events are more difficult to assess.

HCI (keystroke) level: This level consists of input events by the user searching in the DL. The input is usually entered via keyboard, mouse or some other input device such as a braille keyboard. This level corresponds to the keystroke level of the GOMS model[2]

System level: Here events happen on the computer or in the computer network where DL services are executed. This level aggregates very specific information concerning the state of the DL (e.g., database conditions, server load, or amount of network traffic) and its re-sponse (e.g., response time).

3. EVENTS ON THE CONCEPT LEVEL

On the concept level, we have identified several general event types that support comparative evaluation across DLs. Our focus on the concept level represents the centrality of these events for log analysis and interpretation. Events that occur on the concept level indicate critical aspects of the user's interaction with the DL system and supply valuable data for rich interpretation of user behaviour. As is highlighted in other DL logging studies [3], current approaches are often inadequate for capturing complex or abstract actions by the user and are therefore unable to elicit meaningful conclusions. By logging data about general event types at the concept level, we provide a basis for comparative evaluation across DLs.

The event types and event properties that we have identified are neither fixed nor complete and should be viewed as recommendations that can also serve as discussion points in the community. Each event will consist of its own set of properties modelled in XML as sub-elements and attributes of an event. These properties might include, for example a unique *session-id* and *event-id*, but also *timestamps* for start and end of the event, the *service name*, *possible errors* during the event and an indictator of *cancelation* by the user or the system. In addition each event also has an event-specific set of properties.

If it is neccessary to collect more detailed data about a given type of event, we suggest extending standard event definitions through reference to an XML namespace that defines the new properties.

We currently identified the following events on the concept level:

Search A user formulated *query* or *filter condition* that is to be processed by a given DL service against a *collection*.

Navigate A user selects a specific item from a set of possibilities.

Inspect A user accesses the details of a single object.

Display The display event describes a specific *visualization* of the information presented to the user.

Browse A user actions that involve viewing a set of DL objects (e.g., viewing a result list following a search).

Store An object is filed for later reuse, either at a generic location (e.g., a clipboard) or at a specific location (e.g., in a specific folder).

Annotate A user adds information to an existing DL object, either by marking specific parts of it, by linking it to other digital library objects, or by adding inline or external comments.

Authoring A user creates a new DL object or edits an existing object such as a document or annotation.

Help A user request for help information. The help event may be general or context-specific and can include introductory overviews or tutorials about the DL system.

Communicate Users collaborate by communication, which can include posing a simple question or full collaboration through the use of specific tools or services.

In order to analyze the logged events, we have assumed that different stakeholders need different views of the logging data; thus, a variety of analysis tools is required. We have identified the following DL stakeholders: *system owners*, *content providers*, *system administration*, *librarians*, *developers*, *scientific researchers*, and *end-users* of a DL. For each of these stakeholders, different questions should be answered by providing analyse tools.

4. SUMMARY AND OUTLOOK

We have presented the first efforts to develop a standardized experimental framework for digital library evaluation. As an experimental platform for evaluation both within and across DL systems, application of the DAFFODIL framework[3] can substantially advance this research area, since it currently provides functions and services that are based on a solid theoretical foundation and well known models. The proposed logging scheme is a first step intended to encourage evaluation of individual systems as well as comparisons across systems.

Most evaluation techniques require a great deal of preparation and effort and are thus not easily replicated. In case of online DL systems, this means that the results of an evaluation often reflects a past snapshot of the system. It is necessary to find ways for continuous, cost effective and (more or less) automated evaluation of digital libraries. The suggested logging scheme is a first step in this direction. Our group aims at establishing a community forum for evaluators in order to promote the propagation of various tools and approaches, and the exchange of experience.

As part of the effort to encourage a community forum for researchers interested in DL evaluation, we have published documentation for the logging scheme on the forum website at `http://www.is.informatik.uni-duisburg.de/wiki/index.php/JPA_2_-_WP7`. Tools for log analysis, log visualitations and anonymized logging data from two DL services will be made available.

This work was funded by the DELOS Network of Excellence on Digital Libraries (EU 6. FP IST, G038-507618)

5. REFERENCES

[1] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovcs, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Solvberg. Evaluation of digital libraries. (Submitted for publication), 2006.

[2] M. A. Goncalves, R. Shen, E. A. Fox, M. F. Ali, and M. Luo. An xml log standard and tool for digital library logging

[2] http://www.usabilityfirst.com/methods/goms.txl

[3] http://www.dlib.org/dlib/june04/kriewel/06kriewel.html

analysis. In *6th ECDL 2002,Springer. Lect. Notes Comput. Sci. 2458, 129-143*, 2002.

[3] B. Pan. Capturing users behavior in the national science digital library (nsdl). Technical report, NSDL, 2003. `http://dlist.sir.arizona.edu/848/`.