



Preserving the Collective Expressions of the Human Consciousness

Jim Jansen

College of Information Sciences and Technology
The Pennsylvania State University

jjansen@ist.psu.edu



Research and Motivation

- Focus: There should be organized effort to preserve Web search engine transaction logs.
- Motivation: Not preserving the query logs from Web search engines would be (and is) a critical loss of a temporal record of the expression of the collective human consciousness.
- Recommendation: Develop policies, processes, methods, and tools to preserve Web search engine logs.



Search Logs

“ Database of Intentions ”

John Battelle

<http://battellemedia.com/archives/000063.php>

Federated Media/Wired

“ We are looking into the mind of the world.”

Peter Day

<http://news.bbc.co.uk/2/hi/business/4436764.stm>

BBC Radio 4 and BBC World Service





Search Logs

- **Much more.** At a global level, search logs are the collective expressions of the human consciousness. As such, these records deserve preservation.
- **Human consciousness:** the perceived relationship between oneself and one's environment at some temporal point.
- **Queries** are unique in capturing the “gaps” or incidents that trigger one to go to the Web to seek information or services. They provide a unique longitudinal view of these incidents on a scale that exist no where else.
- Search logs contain **expressions of our aspirations, values, character, desires, fears, conflicts, hopes, ...** and anything else that one can express.



Background

- The Web has become an essential facet in the daily lives of many people with impact in nearly every area of human endeavor.
- Search engines are the main portal to the Web. With nearly 70% of Web searchers using a search engine as their point of entry.
- Major Web search engines service millions of queries per day and present billions of results per week [Nielsen 2006]



Background

- These search engines record in transaction logs the interactions among users, the search engine, and Websites.
- The use of data stored in transaction logs of Web search engines, Intranets, and Web sites can provide valuable insight into understanding the information-searching process of online Web searchers.
- Such an understanding can inform information system design, assist interface development, and aid in the devising of information architecture for content collections.



Background

- So far, the search engine companies have received most of the benefits from these query logs.
- The public has access to only the most general and popular of searching statistics. In other words, these transaction logs are a private resource of the search engine companies.
- Conversely, one can view these transaction logs as a public resource. These query logs can and should be viewed as public records that must be preserved as expressions of the collective human consciousness.



Background

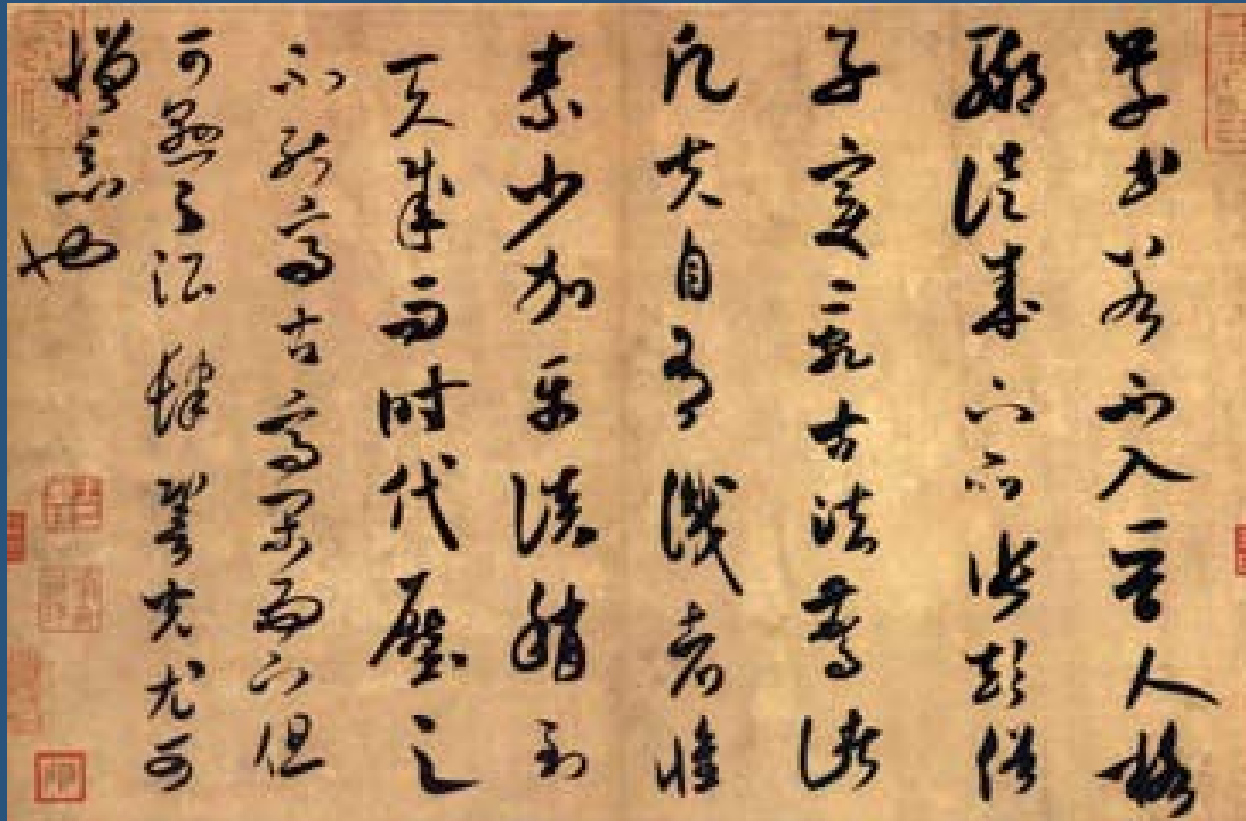
Humans appear to have always found ways to express themselves.



Drawing from Lascaux Cave



Background



There are records from the original Chinese dynasty.



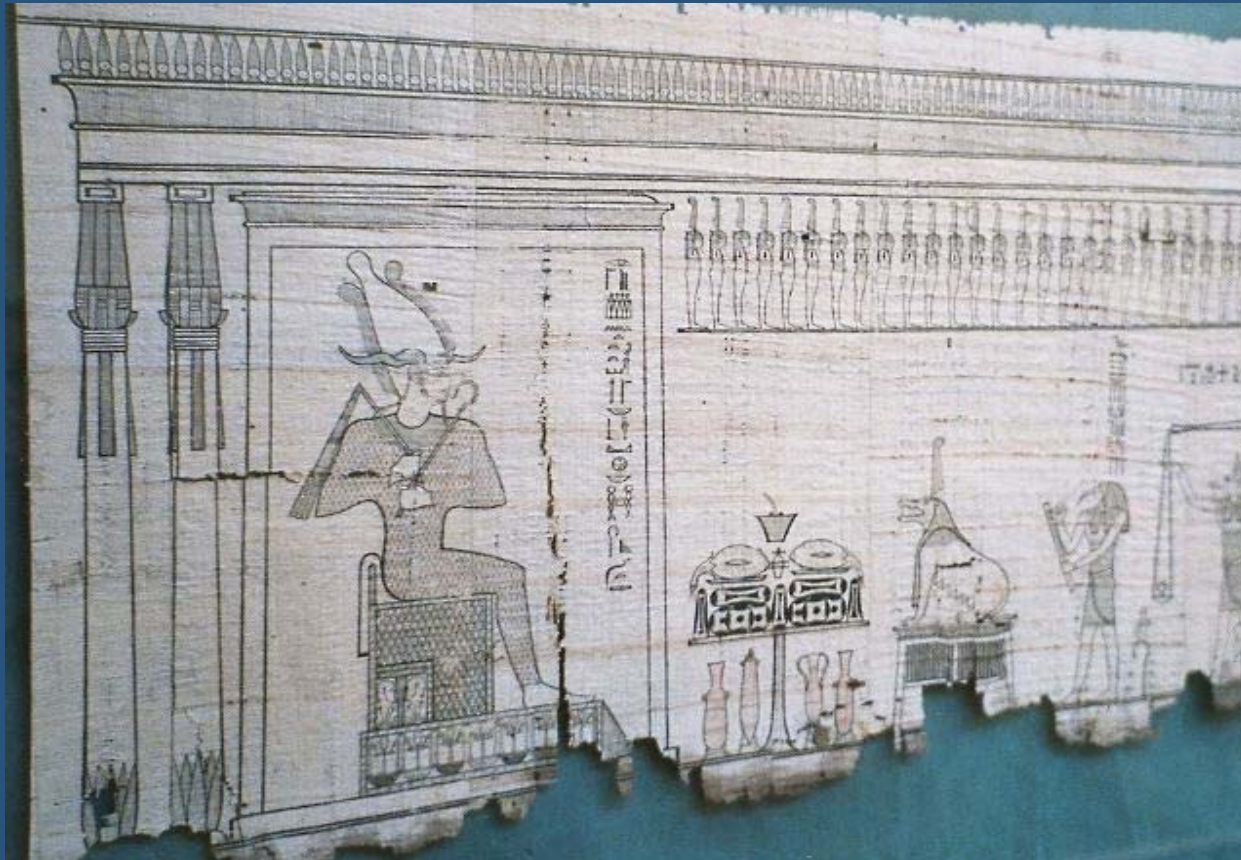
Background



There are rope records from the Inca Empire.



Background



Papyrus scrolls from ancient Egypt.



Background



Pages from the Dead Sea Scrolls.



Background



Original copies of the Gutenberg Bible.



Background

- Interestingly, is that this record of human expression has not carried over into the era of Web search engines.
- We do not know the first search query, even though search engines entered the scene in only the early 1990s.
- We have lost a wealth of vital and important records of human expressions.



Implications

- Preservation of query logs has similarities to other efforts.
- Preserving our electronic heritage was the prime motivator for the Internet Archive Project (a.k.a. the WayBackMachine), which provides access to snapshots of Web pages from about 1996 to the present.
- At the individual level, there is Gordon Bell's MyLifeBits (<http://research.microsoft.com/~gbell/>).
- There should be a similar project for preserving query logs and making these available for researchers, businesses, governments, and the general public.



Proposal

The process of preserving query logs from Web search engines:

1. Begin the collection and preserving of query logs that are currently publicly available.
2. Commence a cooperative partnership among academia, Web search engine companies, and governmental agencies, both national and globally, to establish a more systematic recording and preservation of query logs.
3. Investigate methods for de-identifying log data (e.g., removing transactional information about the user from content information).
4. Develop the architecture for long-term storage, management



Discussion

- Certainly preserving query logs has benefits, risks, and costs.
 - Benefits include preserved records of human expression in an increasingly important area of people's daily lives. One can use these records for a variety of social, commercial, and historical purposes.
 - For risks, we need to do a careful analysis of who is exposed to risk by preserving these records and what are the mitigation approaches, including possibly "shelving" logs for a period of time.
 - Finally, there are costs involved to the owners of the logs. We have to find ways to overcome these costs or provide incentives for preservation.



Questions and Discussion

Jim Jansen

College of Information Sciences and Technology
The Pennsylvania State University

jjansen@ist.psu.edu