OSI and ET: Originating Source of Information and Evidence Traceability

Robert Ball
Center for Human-Computer Interaction
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061, USA
rgb6@vt.edu

Pardha Pyla
Center for Human-Computer Interaction
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061, USA
ppyla@vt.edu

Manuel A. Pérez-Quiñones
Center for Human-Computer Interaction
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061, USA
perez@cs.vt.edu

CHI 2007, April 28 – May 3, 2007, San Jose, USA
ACM 978-1-59593-642-4/07/0004.

Abstract

Originating Source of Information (OSI) is the idea of following all data, facts, and citations that documents rely on for their arguments back their source. OSI then helps people perform Evidence Traceability (ET), which allows them to understand the questions about the different sources used in documents such as how many unique sources were used, where the sources came from, when the sources were obtained, and how the sources were obtained. Answering these questions allows people to better question the validity of documents, reevaluate hypotheses, or continue work or research when original authors are not available.

Keywords

traceability, decision making, satisficing, personal information management

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

The area of Personal Information Management (PIM) deals with the life cycle of creation, dissemination, archival, and accessibility of one’s personal data. Management of Information Systems (MIS) is a field focusing on issues with managing larger, company-level data resources. At some level, both PIM and MIS are often concerned with managing data in the form of documents and solving the problem of where they are located and how they can be accessed.

However, one problem that is not heavily researched is where the data or documents might have come from. Documents are often composed of data from a number of sources. In this work, we address questions such as: What is the original source of the information in this document? Where did it come from? When was it obtained? What was the context in which the information was obtained?

In order to answer these questions we introduce the idea of Originating Source of Information (OSI). OSI is based on the idea that any particular piece of data can be tracked back to its origin if proper meta-data is preserved about that data. As PIM and MIS are concerned with managing documents to keep track of them for later use or reference, OSI is concerned with managing the where, when, and how of all the pieces of the information kept in the documents.

Indeed, OSI concentrates on providing a basis for Evidence Traceability (ET). By "evidence" we mean the data that a document depends on as a basis of argument. By "traceability" we mean that such data can be tracked. Therefore, ET is the tracking of citations, validity, or proof behind arguments made in documents.

Similar to how an attorney provides evidence for or against an argument in a courtroom, ET is the exercise that a person might go through to determine the strength of an argument based on external facts or datum points. ET establishes how many distinct sources or facts that a particular document uses.

As ET is the task of determining the strength of an argument, OSI is a means for better facilitating that task. Simply put, OSI helps manage where, when, and how the "facts" in an argument were obtained to more quickly determine the strength of an argument.

Motivation

In a number of disciplines accurately backing up claims is essential to success. For example, journalism, research, criminology, and intelligence all require that claims and facts be substantiated for documents. Documents that do not cite or reference other sources are often not taken seriously.

However, in disciplines like journalism, time to adequately document, cross reference, and report all sources is not always available due to strict deadlines. As Matthew Arnold [6] clearly expressed, "journalism is literature in a hurry."

For example, consider a group of journalists covering a large story on a corporate scandal. The group works mostly independently of each other giving daily summaries of their work to the editor without necessarily having communicated with each other.

Figure 1 shows how one source of information might be reported as multiple sources of information by different journalists. By receiving information from a variety of different mediums and without keeping careful track of where all the sources and facts originated from could lead the editor, and subsequently the readers, to believe that there are more sources than actually exist.

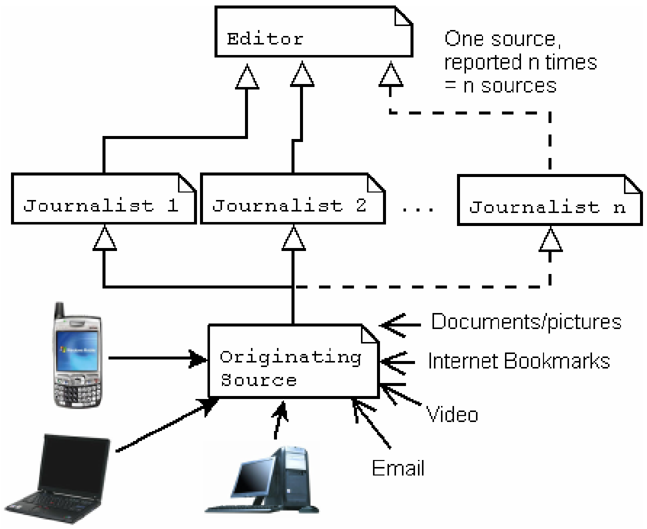

Figure 1. OSI visualization of how one source might be intercepted by a number of different medias and reported as distinct sources.

For instance, Journalist 1 could personally interview an employee about the corporate scandal. Journalist 2, not knowing about the interview could email the same employee for information. Meanwhile, another Journalist might read the employee's blog about the scandal. Although there is only one source, if each journalist did not specifically mention the exact name of the employee in his daily summary of work, the editor might be led to believe that there are three unique sources of information while in fact there is only one.

This can be a serious issue. Heuer [3] points out that the more sources or facts that exist for an argument the more people tend to believe it to be true.

As a result, consider how one incorrect source can lead to a number of wrong conclusions. Journalists often rely on already published articles to check the facts, thus multiplying the strength of a single source.

Although the intelligence community might appear to be different than the field of journalism, Johnston [4] summarizes daily life in the US intelligence community as the following (italics indicate direct quotations from analysts in varying agencies): "*Imagine USA Today with spies - bullet points, short paragraphs, the occasional picture. You know, short and simple*" and "*I think of myself as a writer for the most important newspaper in the world.*" To be brief, Johnston [4] explains that the intelligence community has the same problems of tracing sources and facts as journalists.

Another example is research citations. For instance, take a fictitious research paper that performed a study and concludes that the world is flat. A review paper then cites the research paper as an example of evidence that the world is flat. A third paper then cites both the research paper and the review paper as previous work and evidence that the world is flat. A fourth paper then cites all three papers as evidence and expounds on the idea in a viewport paper which could later be cited as a fourth citation in another paper.

With the above example, only one paper actually provides any actual research claim that the world is flat. However, without meticulously reading all the papers it would *appear* that there are four papers on the subject.

Regardless of discipline, one source can appear to multiply and have a strong influence if not correctly identified. By using ET to track the sources and OSI as a tool to help ET in the effort, sources and facts can be identified.

We hypothesize that knowing the details of the original facts and sources will affect how decisions and policies are made. If the original information is available for future scrutiny, and not just summarized, then decisions and policies can be better made on facts and not just opinions.

Indeed, knowing simple information about the sources of the document, such as when the source was retrieved could change hypotheses or ideas. In a world of constant change knowing the what, when, where, and how (context) of the sources can be invaluable.

Layers

To further motivate the need for OSI consider how most large corporations and organizations work. As a general rule, most large corporations are separated by a number of degrees of management levels. For example, many large corporations often have a CEO (Chief Executive Officer), presidents, vice-presidents, directors, managers, and so on.

Out of necessity, the CEO cannot know about every datum of information that exists in the company. In general, the top executives only have time to review "executive reports" - summaries of reports. Those reports in turn are often summaries of other reports so that the top executives receive summaries based not on facts and figures, but summaries based on averages and other summaries. As a result, with large organizations the most common products produced are paper or electronic reports or summaries [5].

Summaries of work are by definition only reviews or synopsis of the work, not detailed information. As a result many details about sources and facts are lost is the summarized information and not recoverable.

For a simple visual representation of two layers of reporting consider Figure 2. It shows how with only two layers of reporting (where Supervisor 1 creates a report as well) multiple sources or facts can be repeated. For instance, Source X is reported twice to Supervisor 1 by both analysts. Also, both Analyst 2 and Supervisor 1 include Source Y in their reports.

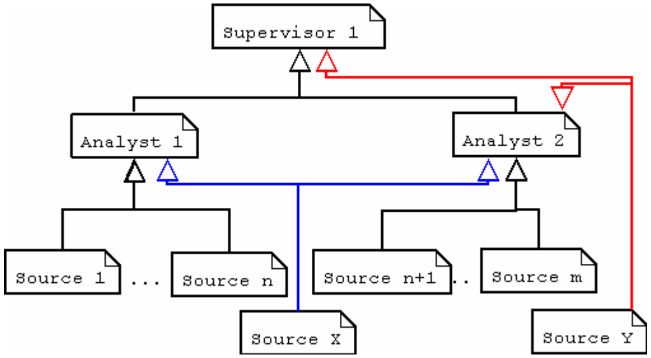

Figure 2. Visual representation of the complexity of two layers of reporting.

Each additional layer adds to the complexity of tracing what sources are included in documents. This gives certain sources or facts greater strength or influential ability than other sources or facts – for example, Sources Y and X both gain greater influential ability over the other sources as they are used in more than one report.

Functionality

Our implementation of OSI is based on a number of ideas such as memex [1] and details-on-demand. In other words, all sources are recorded in such a way that they are available for future deliberation.

By adding meta-data to documents that can be viewed on demand people can refer back to the *original* sources, citations, or facts that make up arguments. Similar to how details-on-demand only shows details when specifically requested, we propose an implementation of OSI that encodes all the information needed in a document without being distracting to the reader that only wants the overview.

As shown in the Figure 2 it is not enough to know that Supervisor 1's report came from Analysts 1 and 2's reports. As a result, an important aspect of our implementation of OSI will have the ability to capture the child sources of the sources cited. Supervisor 1's report would have metadata showing Analyst 1 and 2's reports as sources *and* also have the metadata from those reports available for future reference.

Such metadata can be visualized to quickly ascertain the number of unique sources. Automatic algorithms could also be engaged to understand basic statistics about the number of sources, the homogeneity of sources came from, etc.

Approaches

There are a number of different approaches that one can take to implementing OSI. We explain a best-try approach and a process-oriented approach in this section.

Best-try Approach

The best-try approach takes the idea that "anything is better than nothing." Without any systematic process to rely on, the authors of documents add metadata to their documents as best they can.

For example, a journalist might include simple metadata of what people and organizations he has communicated with to create an article. The metadata might include the time, date, and contact information of the different people.

The results of this best-try approach could allow other journalists to revisit the same issue or topic with the same contact information that the original journalist had even if the original journalist is unavailable (e.g. retired, left the company, sick, etc.). This effectively allows documents to have a life of their own without being tied down to a single person.

A similar approach has been provided by Oculus's Sandbox [7]. Sandbox is an analytical task environment whose state can be stored and recorded as metadata in documents. For example, if a group of scientists had performed brainstorming sessions, each session can be recorded as metadata so that others can refer back and examine the analytical process of their thinking.

Process-Oriented Approach

As opposed to the best-try approach where as much data is recorded as deemed necessary and appropriate, the process approach records each source as a separate piece of data with a unique identification number. This approach has the advantage that documents contain smaller amounts of metadata, but has the disadvantage that in order to understand what the metadata means access to the repository is required.

Citeseer [2] is an example of such a repository. If a researcher cites a document contained in the Citeseer repository then all the documents that that document cites and the documents those documents cite can be traced (at least theoretically).

For non-research organizations, such as the intelligence and journalism this would require a new process, which might be better, but less likely to catch on. As Abigail Sellen and Richard Harper (IEEE Award Winners) explain, a different managerial process is needed for new technologies to be fully utilized [5].

Conclusion

Talented writers are able to manipulate people's feelings and thoughts with little or no facts or sources. Viewpoints and policies can be changed or encourage by persuasive literature.

However, by revealing the original facts and sources that underscore arguments then more objective thinking can be performed. By performing ET with the help of OSI the following can occur:

- Hypotheses and claims can be directly challenged by going back to the original sources of data or facts.
- Decision-making and satisficing can be influenced by the ability to get to the original information to reinterpret and reevaluate a given situation instead of relying on another's opinion or summary.
- Follow-up ideas or articles can continue even if the original author is no longer available.

We are currently in the early phases of creating an implementation of OSI. We will then rigorously test how decisions are made with ET and OSI.

Acknowledgements

We thank Don Jones from Pacific North-West National Laboratories (PNNL) for his help and ideas.

REFERENCES

[1] Bush, Vannevar. As we may think. *Atlantic Monthly*, July 1945.

[2] Citeseer. http://citeseer.ist.psu.edu/

[3] Heuer, Richard, Psychology of Intelligence Analysis, *Center for the Study of Intelligence*, 1999. ISBN 1594546797.

[4] Johnston, R., Analytic Culture in the U.S. Intelligence Community, Center for the Study of Intelligence, *Government Printing Office,* Pittsburgh, PA*.* ISBN 1-929667-13-2.

[5] Sellen, Abigail and Harper, Richard. The myth of the paperless office. *The MIT Press*. 2002. ISBN 0-262-19464-3.

[6] Smelstor, Majorie and Weiher, Carol. Using Popular Culture to Teach Composition. *The English Journal*, March 1976, p. 41-46.

[7] Wright, W., Schroh, D., Proulx, P., Skaburskis, A., and Cort, B. The sandbox for analysis – concepts and methods. *Proceedings of CHI 2006*, 801-810.