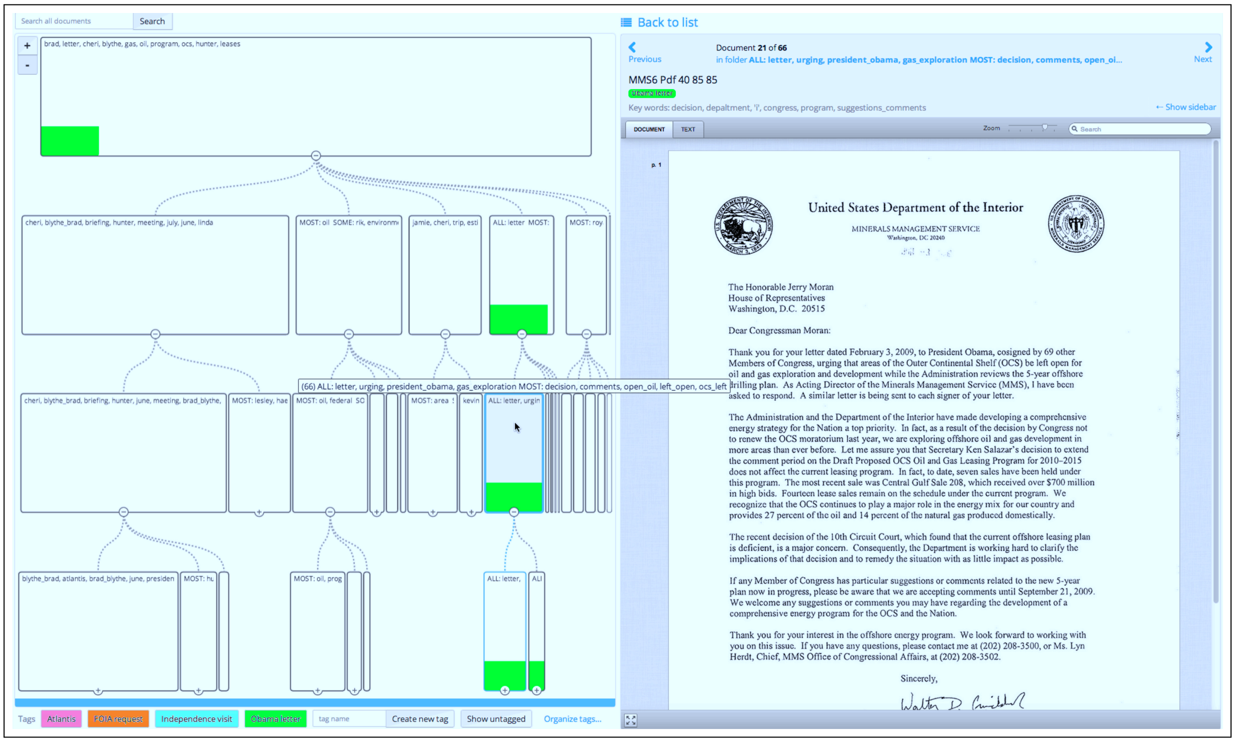
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2346431, IEEE Transactions on Visualization and Computer Graphics

Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists

Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner, *Member, IEEE*


Search all documents
Search
+
-
brad, letter, cheri, blythe, gas, oil, program, ocs, hunter, leases
cheri, blythe_brad, briefing, hunter, meeting, july, june, linda
MOST: oil SOME: rik, environm
jamie, cheri, trip, esti
ALL: letter MOST:
MOST: roy
(66) ALL: letter, urging, president_obama, gas_exploration MOST: decision, comments, open_oil, left_open, ocs_left
cheri, blythe_brad, briefing, hunter, june, meeting, brad_blythe,
MOST: lesley, hae
MOST: oil, federal SO
MOST: area
kevin
ALL: letter, urgin
blythe_brad, atlantis, brad_blythe, june, presiden
MOST: hu
MOST: oil, prog
ALL: letter,
ALI
Tags
Atlantis
FOIA request
Independence visit
Obama letter
tag name
Create new tag
Show untagged
Organize tags...
Back to list
Previous
Document 21 of 66
in folder ALL: letter, urging, president_obama, gas_exploration MOST: decision, comments, open_oi...
Next
MMS6 Pdf 40 85 85
Obama letter
Key words: decision, depaltment, 'i', congress, program, suggestions_comments
← Show sidebar
DOCUMENT
TEXT
Zoom
Search
p. 1

United States Department of the Interior

MINERALS MANAGEMENT SERVICE
Washington, DC 20240

The Honorable Jerry Moran
House of Representatives
Washington, D.C. 20515

Dear Congressman Moran:

Thank you for your letter dated February 3, 2009, to President Obama, cosigned by 69 other Members of Congress, urging that areas of the Outer Continental Shelf (OCS) be left open for oil and gas exploration and development while the Administration reviews the 5-year offshore drilling plan. As Acting Director of the Minerals Management Service (MMS), I have been asked to respond. A similar letter is being sent to each signer of your letter.

The Administration and the Department of the Interior have made developing a comprehensive energy strategy for the Nation a top priority. In fact, as a result of the decision by Congress not to renew the OCS moratorium last year, we are exploring offshore oil and gas development in more areas than ever before. Let me assure you that Secretary Ken Salazar's decision to extend the comment period on the Draft Proposed OCS Oil and Gas Leasing Program for 2010–2015 does not affect the current leasing program. In fact, to date, seven sales have been held under this program. The most recent sale was Central Gulf Sale 208, which received over $700 million in high bids. Fourteen lease sales remain on the schedule under the current program. We recognize that the OCS continues to play a major role in the energy mix for our country and provides 27 percent of the oil and 14 percent of the natural gas produced domestically.

The recent decision of the 10th Circuit Court, which found that the current offshore leasing plan is deficient, is a major concern. Consequently, the Department is working hard to clarify the implications of that decision and to remedy the situation with as little impact as possible.

If any Member of Congress has particular suggestions or comments related to the new 5-year plan now in progress, please be aware that we are accepting comments until September 21, 2009. We welcome any suggestions or comments you may have regarding the development of a comprehensive energy program for the OCS and the Nation.

Thank you for your interest in the offshore energy program. We look forward to working with you on this issue. If you have any questions, please contact me at (202) 208-3500, or Ms. Lyn Herdt, Chief, MMS Office of Congressional Affairs, at (202) 208-3502.

Sincerely,

Walter D. Cruickshank

Fig. 1. *Overview* is a multiple-view application intended for the systematic search, summarization, annotation, and reading of a large collection of text documents, hierarchically clustered based on content similarity and visualized as a tree (left). Pictured: a collection of White House email messages concerning drilling in the Gulf of Mexico prior to the 2010 BP oil spill.

Abstract—For an investigative journalist, a large collection of documents obtained from a Freedom of Information Act request or a leak is both a blessing and a curse: such material may contain multiple newsworthy stories, but it can be difficult and time consuming to find relevant documents. Standard text search is useful, but even if the search target is known it may not be possible to formulate an effective query. In addition, summarization is an important non-search task. We present *Overview*, an application for the systematic analysis of large document collections based on document clustering, visualization, and tagging. This work contributes to the small set of design studies which evaluate a visualization system "in the wild", and we report on six case studies where *Overview* was voluntarily used by self-initiated journalists to produce published stories. We find that the frequently-used language of "exploring" a document collection is both too vague and too narrow to capture how journalists actually used our application. Our iterative process, including multiple rounds of deployment and observations of real world usage, led to a much more specific characterization of tasks. We analyze and justify the visual encoding and interaction techniques used in *Overview*'s design with respect to our final task abstractions, and propose generalizable lessons for visualization design methodology.

Index Terms—Design study, investigative journalism, task and requirements analysis, text and document data, text analysis.

1 Introduction

Freedom of Information Act (FOIA) requests, leaks, government transparency initiatives, or other disclosures can result in thousands or millions of pages of potentially newsworthy material. Investigative journalists must find the stories lurking in these massive document collections, but it is frequently impossible to read every document. Standard text search can be used to locate documents containing particular terms, but not all information retrieval problems can be expressed as word search queries, especially if the relevant information is unexpected or novel. Journalists may also be interested in patterns of text across many documents, which can reveal significant trends, categories, or themes. We conjectured that this *document mining* problem could be solved by a visualization system built around clustering and tagging documents. The path from this hypothesis to a system that working journalists would voluntarily use was a long one; we needed to refine both our understanding of the problem and the ways in which journalists might want to solve it.

This paper reports on the design, adoption, and analysis of

- *Matthew Brehmer, Stephen Ingram, and Tamara Munzner are with the University of British Columbia. E-mail: {brehmer,sfingram,tmm}@cs.ubc.ca.*
- *Jonathan Stray is with the Columbia Journalism School and the Associated Press. E-mail: jonathanstray@gmail.com.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication 9 Nov. 2014; date of current version 1 Aug. 2014. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

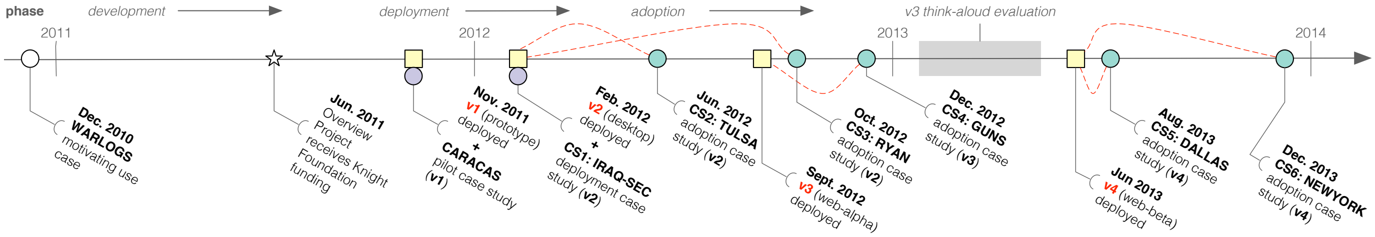

phase
development
deployment
adoption
v3 think-aloud evaluation
2011
2012
2013
2014
Dec. 2010 WARLOGS motivating use case
Jun. 2011 Overview Project receives Knight Foundation funding
Nov. 2011 v1 (prototype) deployed + CARACAS pilot case study (v1)
Feb. 2012 v2 (desktop) deployed + CS1: IRAQ-SEC deployment case study (v2)
Jun. 2012 CS2: TULSA adoption case study (v2)
Sept. 2012 v3 (web-alpha) deployed
Oct. 2012 CS3: RYAN adoption case study (v2)
Dec. 2012 CS4: GUNS adoption case study (v3)
Jun 2013 v4 (web-beta) deployed
Aug. 2013 CS5: DALLAS adoption case study (v4)
Dec. 2013 CS6: NEWYORK adoption case study (v4)

Fig. 2. Timeline of *Overview*'s development, deployment, and adoption phases: deployments are represented as yellow squares; deployment-phase case studies are represented as purple circles, while adoption-phase case studies are represented as green circles. The dotted red lines indicate which version of *Overview* was used in each case study.

Overview (http://overviewproject.org), an application developed by the Associated Press in collaboration with our research group over several years. *Overview*, shown in Figure 1, visualizes a document collection as a tree where nodes represent clusters of similar documents; users can navigate this tree, identify clusters, read individual documents, and annotate documents with meaningful tags. A timeline illustrating *Overview*'s development, deployment, and adoption phases is shown in Figure 2. Beginning with a motivating use case, we produced a research prototype (*v1*), developed a publicly available cross-platform desktop application (*v2*), and finally a web-based application (*v3-v4*). Ultimately, we succeeded in building a useful tool for journalists: we report on multiple case studies where *Overview* was adopted for real investigations. Analysis of these cases revealed that journalists often used the application in ways we did not anticipate, and we found that the often-used concept of "exploring" a document collection fails to capture the tasks that journalists actually perform.

Contributions: We frame this work as a visualization design study, a process of iterative design and evaluation addressing a particular domain problem, involving collaborators and users from that domain [46]. The contributions of this paper include *Overview* itself, our characterization of data and task abstractions, a description of its usage in real investigations spanning four deployments and six case studies, and a detailed analysis of the mapping from these abstractions to visual encoding and interaction design choices. This analysis led to important design revisions, based on a better understanding of *why* and *how* journalists use *Overview*. From this experience we propose generalizable lessons for visualization design methodology.

Outline: We begin with a survey of related work in Section 2. We then describe our initial motivating use case in Section 3. The design of *Overview* is presented in Section 4, which includes our initial task abstraction, *Overview*'s underlying data abstractions, and a description of its user interface. In Section 5, we report on real world usage of *Overview* by six journalists who used it for their own investigations; in five of these cases, the investigation resulted in a published story. Based on our observations of what these users did, we revisit our initial task abstraction and reflect upon the rationale for *Overview*'s visual encoding and interaction design choices in Section 6. Finally, in Section 7, we reflect on the methodological implications of our approach, and Section 8 summarizes our contributions.

2 Related Work

There have been a number of approaches and tools to support the analysis of document collections, spanning a range of data transformations and visual encodings. We also review how these tools were evaluated.

Topic model visualizations: One common approach to visualizing a document collection uses probabilistic topic models inferred from the collection. These define topics as distributions of words and assign a distribution of topics per document. Both distributions are visualized directly in recent work by Chaney and Blei [6], while other systems focus on the number of documents in each topic [9, 10, 34], or use the topic assignments to compute similarity for document-based visualizations [7, 11]. *Overview* does not use distribution-based topic models but directly creates a hard hierarchical clustering, which is presented in a document-based tree visualization.

Documents as points: Many systems, including the first two versions of *Overview*, encode individual documents as points in a scatterplot. *InfoSky* [18] places points according to a pre-existing hierarchical arrangement of documents; in contrast, *Overview* is intended for document collections without pre-existing hierarchical structure. Other approaches begin with an unstructured document collection and place points based on document similarity metrics and dimensionality reduction techniques, such as *Leaksplorer* [2], *PEx* [40], and *EV* [7]. *Overview v1-v2* included a similar scatterplot which placed points by dimensionality reduction through multidimensional scaling. Finally, *ForceSPIRE* [12] and *TopicViz* [11] incorporate a scatterplot where document-points can be interactively placed according to the user's own semantics, adaptively adjusting the underlying similarity metric used between document pairs. In Section 6.2, we discuss in greater detail why a scatterplot was omitted from later versions of *Overview*, and how tagging documents and clusters is an effective alternative to interactive placement.

Documents as landscapes or clouds: Document collections have also been encoded as landscapes, three-dimensional representations of two-dimensional scatterplots where height represents density, as in *In-Spire* [23] and recent work by Österling et al. [39]. However, empirical studies have shown that spatial landscapes are not well suited for encoding inherently non-spatial data, and exhibit poor visual memory performance in comparison to two-dimensional scatterplots [51].

It is also possible to visualize a document collection by encoding clusters of documents as interactive tag clouds, as in *Newdle* [35]. Once again, previous research has documented the perceptual drawbacks of tag clouds [22]. By encoding a document collection as a hierarchical tree, *Overview* circumvents these issues.

Documents as networks of entities: *Jigsaw*'s approach [16, 28] to document collection analysis differs from *Overview* in that it emphasizes the extraction of entities from documents, linking names, places, events, and dates, constructing visualizations around these relationships. The emphasis on entities is reflective of the domains in which *Jigsaw* is used, which include intelligence analysis, law enforcement, and academic research [28]. Journalists frequently start with barely-legible scanned documents which must first be converted to text through Optical Character Recognition (OCR), greatly reducing the accuracy of standard entity extraction techniques. As a flexible multiple-view application, *Jigsaw* also has a significant learning curve, and users have reported investing many months into learning and using it [28]. The journalists we spoke to are accustomed to short deadlines and may only intermittently be working on a story involving a large document collection, so simplicity is a crucial feature.

Documents as trees and rivers: Like *Overview*, the *HierarchicalTopics* system [10] features a hierarchical tree visualization of document clusters, initially arranged by similar keywords. It allows users to re-arrange the tree according to their own semantics, similar to how *ForceSPIRE* users can rearrange documents in a scatterplot [12]. *HierarchicalTopics* [10] additionally allow users to track topic prevalence over time with a *ThemeRiver* visualization [21]. However, this approach requires temporal metadata that would be difficult to extract from the diverse document sources supported by *Overview*.

Evaluating visual document mining tools: Several of the aforementioned tools have been evaluated via controlled experiments and case studies. Controlled experiments, such as those used to evaluate *Newdle* [35] or *HierarchicalTopics* [10], often involve non-specialist users conducting domain-agnostic tasks specified by the researchers, who conjecture that they match with real world usage. Moreover, the documents used in these controlled experiments were collections of online news articles which are not appropriate test data for *Overview*, as professionally produced news articles are clean and homogeneous, unlike the diverse and messy documents obtained by our case study journalists, which often contain little or no metadata; news articles are the *output* of the journalistic document mining process, not the input.

Most similar to our approach is a series of case studies of academic researchers, intelligence analysts, and law enforcement personnel who had adopted *Jigsaw* [28]. These case studies resulted in a better understanding of *Jigsaw*'s utility in relation to users' domain-specific tasks; like us, they identified similar barriers to adoption and their results suggested new directions for design [16, 17].

3 Motivating Use Case

The *Overview* project began in December 2010, when Associated Press journalist and co-author Stray visualized a subset (11,616 of 391,832) of the WikiLeaks Iraq War Logs [49]. Journalists had previously examined these documents by using text search to retrieve specific records and by visualizing the structured data fields such as time and location, but had not attempted analysis of the bulk unstructured text of the reports. In this visualization, which we will refer to as WARLOGS, documents were represented as points placed according to a measure of similarity between documents and coloured according to pre-existing categorical labels, such as *"friendly action"* and *"criminal incident."* As shown in Figure 3, this technique revealed meaningful cluster structure that cross-cuts the colourings, showing that the pre-existing coarse categorization does not capture the whole story.

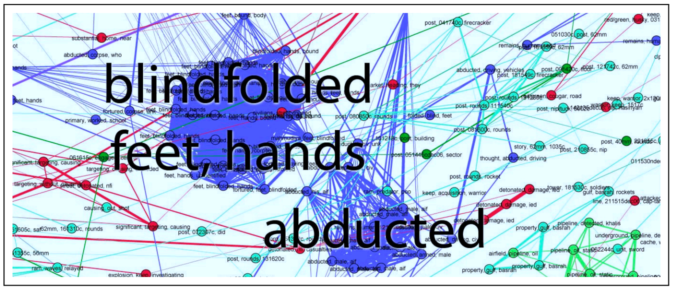


Fig. 3. Detail from *"A full-text visualization of the Iraq War Logs"* (WARLOGS) [49], in which distinct clusters of documents are visible; these documents pertain to "criminal incidents" during the Iraqi civil war involving abductions and blindfolding.

The WARLOGS visualization had serious limitations: it was not possible to interactively and systematically examine the contents of clusters of documents. However, it demonstrated that visual cluster analysis could illuminate previously unknown and meaningful structure in a real world document collection, a conjecture that Stray had synthesized from his previous experience reporting on this collection of documents. On the basis of this promising result, Stray collaborated with us to design an interactive visualization tool for document mining.

4 Overview Design

We now summarize our initial task abstraction, *Overview*'s underlying data abstractions, and the elements of its user interface.

Initial task abstraction: During the development of *Overview v1-v2* our task abstraction was based on the WARLOGS use case: journalists would be motivated by the hypothesis that their document collection contained a semantically interesting cluster structure, and would require a means for *exploring* that structure, drilling down into these clusters to examine the contained documents. During this exploration, they would need a way to keep track of what they had discovered, allowing them to revisit previously examined clusters and documents.

Data abstractions: Although *Overview*'s design has evolved over the course of four deployed versions, it continues to reflect several underlying data abstractions. *Overview* does not incorporate any novel text analysis techniques; following a practice common in that domain, we convert each document to a vector of words weighted by the Term Frequency–Inverse Document Frequency (TF-IDF) formula, and compute similarity between documents using the cosine distance metric [43]. We generate our document clusters by hierarchically clustering these distances and encoding the result as a tree [24, 25]. Clusters are labeled with keywords extracted via TF-IDF scores.

Multiple meaningful clusterings may exist for any collection of documents [19]; our particular distance metric and hierarchical clustering algorithm is but one possible choice. User-generated clusterings that leverage domain knowledge can complement automatic clusterings [10, 12]. For these reasons, *Overview* allows an arbitrary number of user-defined *tags* on each document, which can be assigned individually or at the cluster level. Tags allow users to keep track of what they have found and where they have looked so far.

User interface: With each deployment came changes to the user interface, though we will focus on the differences between *Overview v2* and *v4*, shown in Figure 4 and 5, respectively. The visualization design of *v1* and *v2* are quite similar to each other, as are *v3* and *v4*.

Common to all deployed versions of *Overview* is the *Topic Tree* visualization, representing a hierarchical clustering of similar documents, the *Document List*, showing currently selected documents, the *Document Viewer*, and the ability to create and assign custom categorical tags to clusters or individual documents; tags are visually encoded as coloured labels on documents and clusters. Selections of documents are propagated and highlighted across views.

The *Topic Tree* underwent some of the most significant changes. It was redesigned to emphasize nodes, and to visually encode the number of documents in each node, instead of focusing on the edges between identically-sized nodes. In *v1-v2*, the *Topic Tree* could be pruned based on a threshold cluster size, controlled using a set of coloured radio buttons below; in *v3*, we replaced threshold pruning with an open/close interface that allows the user to show or hide the children of any node. Pan and zoom controls were also added, including an auto-zoom feature that automatically zooms and pans to a selected node.

Another prominent change was the removal of the interactive scatterplot visualization, in which individual documents were encoded by points and their placement corresponded to a two-dimensional projection of the original high-dimensional TF-IDF vector space, generated via multidimensional scaling; pairs of documents appearing closer together were deemed to be more similar than pairs of documents that were farther apart. The scatterplot had panning and zooming controls, and document-points could be selected via clicking or lassoing.

We also removed the *Cluster List* and consolidated the *Document Viewer* with the *Document List* (cf. Figure 1). The *Document List* now displays the document title, extracted keywords, and coloured labels indicating which tags have been applied to each document. We added full-text keyword search in *v4*; documents matching a search query are highlighted with colour labels in the *Topic Tree*, and these results can be saved as a persistent tag. Finally, we added a *"Show Untagged"* button in *v4*, which highlights documents and clusters where no tags have been applied, a crucial feature for the (initially unexpected) task of exhaustively reviewing a document collection.

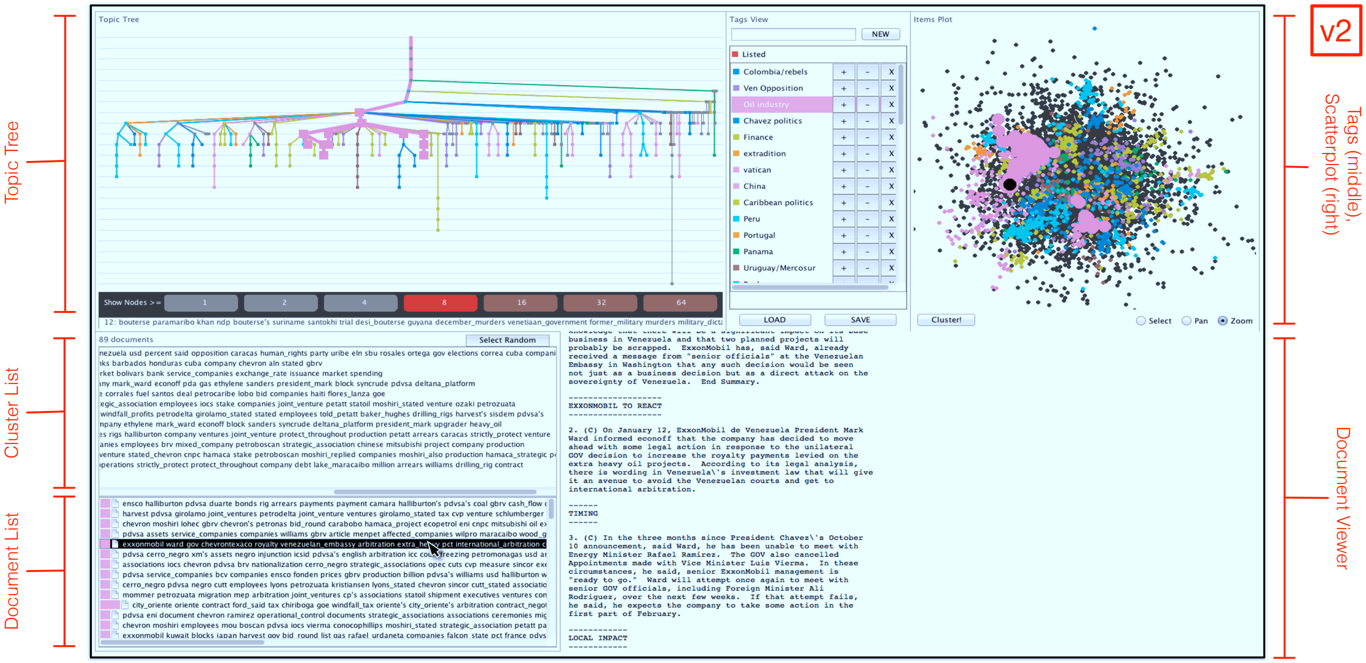
This section summarizes the design without providing any rationale for its evolution. Our decisions were based on observations of real world usage; we provide concrete examples of why and how *Overview* was used by journalists in Section 5. Then, in Section 6, we present our final task abstraction, the outcome of analyzing these observations, and justify our design choices with respect to these revisited tasks.

5 Observations of Real World Usage

We conducted six in-depth case studies where we analyzed the use of *Overview* by investigative journalists. We distinguish between a *case study* and a *usage scenario* [46], in which the former involves a target domain user who uses a tool to examine their own data, having goals related to their ongoing work; in contrast, the latter reports usage of a tool by its designers with curated data and conjectured tasks.

Pilot case study (**CARACAS**): The first user of *Overview* was the Associated Press Caracas bureau chief, whom we asked in November 2011 to use the *v1* prototype to examine 6,849 of the 251,287 U.S. State Department diplomatic cables released by WikiLeaks, those per-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2346431, IEEE Transactions on Visualization and Computer Graphics



12: bouterse paramaribo khan ndp bouterse's suriname santokhi trial desi_bouterse guyana december_murders venetiaan_government former_military murders military_dicta

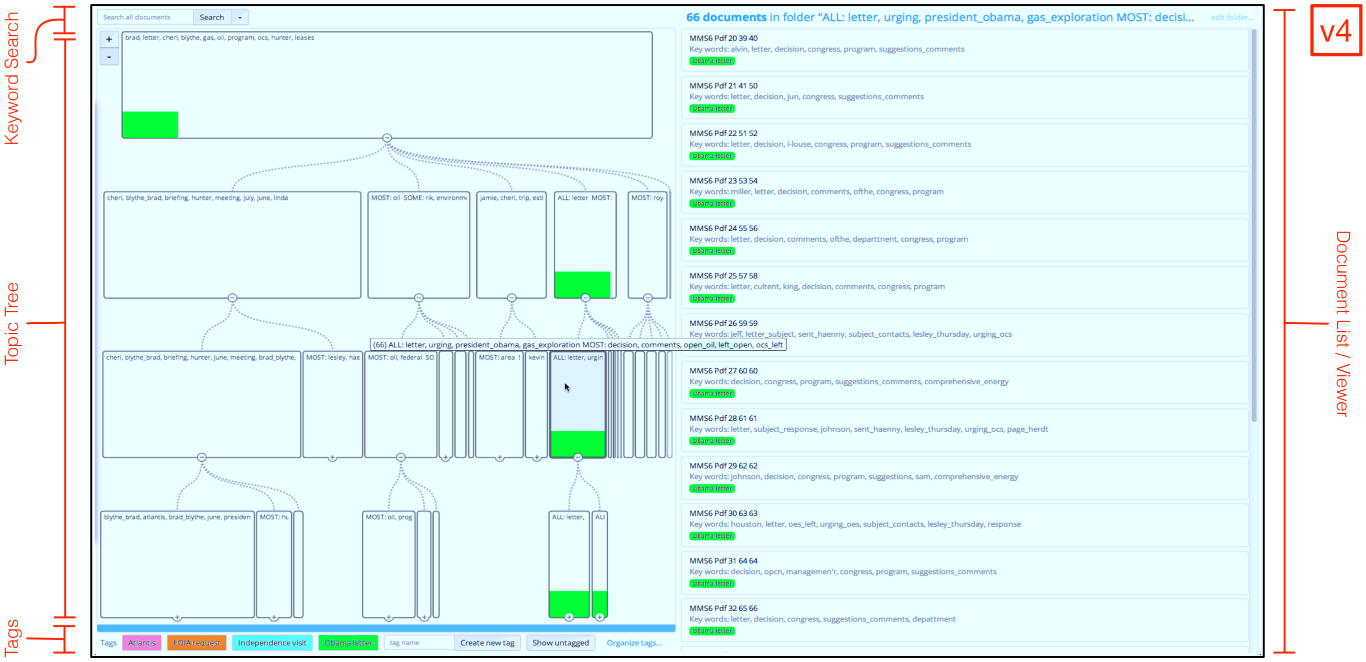
89 documents

```
nezuela usd percent said opposition caracas human_rights party uribe eln sbu rosales ortega gov elections correa cuba compani
ıks barbados honduras cuba company chevron aln stated gbrv
rket bolivars bank service_companies exchange_rate issuance market spending
ıny mark_ward econoff pda gas ethylene sanders president_mark block syncrude pdvsa deltana_platform
corrales fuel santos deal petrocaribe lobo bid companies haiti flores_lanza goe
egic_association employees iocs stake companies joint_venture petatt statoil moshiri_stated venture ozaki petrozuata
windfall_profits petrodelta girolamo_stated stated employees told_petatt baker_hughes drilling_rigs harvest's sisdem pdvsa's
npany ethylene mark_ward econoff block sanders syncrude deltana_platform president_mark upgrader heavy_oil
es rigs halliburton company ventures joint_venture protect_throughout production petatt arrears caracas strictly_protect venture
anies employees brv mixed_company petroboscan strategic_association chinese mitsubishi project company production
venture stated_chevron cnpc hamaca stake petroboscan moshiri_replied companies moshiri_also production hamaca_strategic p
perations strictly_protect protect_throughout company debt lake_maracaibo million arrears williams drilling_rig contract
```

```
ensco halliburton pdvsa duarte bonds rig arrears payments payment camara halliburton's pdvsa's coal gbrv cash_flow
harvest pdvsa girolamo joint_ventures petrodelta joint_venture ventures girolamo_stated tax cvp venture schlumberger
chevron moshiri lohec gbrv chevron's petronas bid_round carabobo hamaca_project ecopetrol eni cnpc mitsubishi oil ex
pdvsa assets service_companies companies williams gbrv article menpet affected_companies wilpro maracaibo wood_g
exxonmobil ward gov chevrontexaco royalty venezuelan_embassy arbitration extra_heavy pct international_arbitration c
pdvsa cerro_negro xm's assets negro injunction icsid pdvsa's english arbitration icc cou freezing petromonagas usd ar
associations iocs chevron pdvsa brv nationalization cerro_negro strategic_associations opec cuts cvp measure sincor exe
pdvsa service_companies bcv companies ensco fonden prices gbrv production billion pdvsa's williams usd halliburton w
cerro_negro pdvsa negro cutt employees lyons petrozuata kristiansen lyons_stated chevron sincor cutt_stated associatio
mommer petrozuata migration mep arbitration joint_ventures cp's associations statoil shipment executives ventures con
city_oriente oriente contract ford_said tax chiriboga goe windfall_tax oriente's city_oriente's arbitration contract_negot
pdvsa eni document chevron ramirez operational_control documents strategic_associations associations ceremonies mig
chevron moshiri employees mou boscan pdvsa iocs vierma conocophillips moshiri_stated strategic_association petatt pa
exxonmobil kuwait blocks japan harvest gov bid_round list gas rafael urdaneta companies falcon state pct france pdvs
```

```
business in Venezuela and that two planned projects will
probably be scrapped.  ExxonMobil has, said Ward, already
received a message from "senior officials" at the Venezuelan
Embassy in Washington that any such decision would be seen
not just as a business decision but as a direct attack on the
sovereignty of Venezuela.  End Summary.

-------------------
EXXONMOBIL TO REACT
-------------------

2. (C) On January 12, ExxonMobil de Venezuela President Mark
Ward informed econoff that the company has decided to move
ahead with some legal action in response to the unilateral
GOV decision to increase the royalty payments levied on the
extra heavy oil projects.  According to its legal analysis,
there is wording in Venezuela\'s investment law that will give
it an avenue to avoid the Venezuelan courts and get to
international arbitration.

------
TIMING
------

3. (C) In the three months since President Chavez\'s October
10 announcement, said Ward, he has been unable to meet with
Energy Minister Rafael Ramirez.  The GOV also cancelled
Appointments made with Vice Minister Luis Vierma.  In these
circumstances, he said, senior ExxonMobil management is
"ready to go."  Ward will attempt once again to meet with
senior GOV officials, including Foreign Minister Ali
Rodriguez, over the next few weeks.  If that attempt fails,
he said, he expects the company to take some action in the
first part of February.

------------
LOCAL IMPACT
------------
```

Fig. 4. *Overview v2*, a desktop application released in Winter 2012. Shown here is 6,849 of the U.S. State Department diplomatic cables released by WikiLeaks, those pertaining to Venezuela. The *"Oil industry"* tag is selected; clusters containing documents having this tag are emphasized in pink in the *Topic Tree* and are shown in the *Cluster List* as a set of keywords. Individual documents having the *"Oil industry"* tag are emphasized in the scatterplot and shown in the *Document List* as a set of keywords. The fifth document is selected; its contents are displayed in the *Document Viewer* and it is marked as a larger black dot in the scatterplot.



(66) ALL: letter, urging, president_obama, gas_exploration MOST: decision, comments, open_oil, left_open, ocs_left

66 documents in folder "ALL: letter, urging, president_obama, gas_exploration MOST: decisi...

MMS6 Pdf 20 39 40
Key words: alvin, letter, decision, congress, program, suggestions_comments
Obama letter

MMS6 Pdf 21 41 50
Key words: letter, decision, jun, congress, suggestions_comments
Obama letter

MMS6 Pdf 22 51 52
Key words: letter, decision, i-louse, congress, program, suggestions_comments
Obama letter

MMS6 Pdf 23 53 54
Key words: miller, letter, decision, comments, ofthe, congress, program
Obama letter

MMS6 Pdf 24 55 56
Key words: letter, decision, comments, ofthe, departtnent, congress, program
Obama letter

MMS6 Pdf 25 57 58
Key words: letter, cultent, king, decision, comments, congress, program
Obama letter

MMS6 Pdf 26 59 59
Key words: jeff, letter_subject, sent_haenny, subject_contacts, lesley_thursday, urging_ocs

MMS6 Pdf 27 60 60
Key words: decision, congress, program, suggestions_comments, comprehensive_energy
Obama letter

MMS6 Pdf 28 61 61
Key words: letter, subject_response, johnson, sent_haenny, lesley_thursday, urging_ocs, page_herdt
Obama letter

MMS6 Pdf 29 62 62
Key words: johnson, decision, congress, program, suggestions, sam, comprehensive_energy
Obama letter

MMS6 Pdf 30 63 63
Key words: houston, letter, oes_left, urging_oes, subject_contacts, lesley_thursday, response
Obama letter

MMS6 Pdf 31 64 64
Key words: decision, opcn, managemen'r, congress, program, suggestions_comments
Obama letter

MMS6 Pdf 32 65 66
Key words: letter, decision, congress, suggestions_comments, depattment
Obama letter

Fig. 5. *Overview v4*, a web-based application released in Summer 2013. Shown here is 625 White House email messages concerning drilling in the Gulf of Mexico prior to the 2010 BP oil spill. The *"Obama letter"* tag is selected; clusters containing documents having this tag are highlighted in green in the *Topic Tree*. One of these clusters is selected and its keywords are displayed in a tooltip; the 66 documents in this cluster are listed in the *Document List*. Selecting a document from this list reveals the *Document Viewer* (cf. Figure 1).

taining to Venezuela; this document collection is featured in Figure 4. Although he found the system interesting, his analysis did not lead to a published story. This informal pilot case study revealed basic usability problems and the experience prompted us to formalize the case study process and identify foci of interest, such as utility, usability, learnability, and journalists' tasks in context.

Metrics: In addition to the qualitative analysis of journalists' tasks, we also focus on the metric of adoption defined as *self-initiated* use: did a journalist freely chose to use the tool for their own investigation, rather than trying out the tool in response to direct solicitation by the researchers? According to this distinction, adoption occurred in five of the six case studies we report, as indicated by the green circles in Figure 2; the journalist in the remaining case study (IRAQ-SEC) was co-author Stray. We were also interested in the outcome of a journalist's investigation: did they complete their investigation to satisfaction as a result of using *Overview*, either by choosing to publish a story or by deciding that their findings did not merit a story? Or did they abandon *Overview* because the tool did not help their investigation?

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2346431, IEEE Transactions on Visualization and Computer Graphics

Recruitment: Since the *v2* deployment, Stray has promoted *Overview* within the data journalism community. Several hundred journalists have created accounts on the public server, and they have collectively uploaded more than 9 million documents; *Overview* is used by approximately 200 unique users each month. At the time of writing we know of nine published stories where *Overview* played a part in the investigative process [3], five of which are discussed as case studies below. The self-initiated journalists featured in case studies 2-6 were recruited after they contacted Stray with technical questions, which often pertained to workflow difficulties such as wrangling their document collection into a format that *Overview* could ingest.

Methods: Our case study findings are the result of triangulating between multiple data collection and analysis methods. Our primary data collection method was that of a semi-structured interview. We conducted interviews via Skype or Google+ Hangout, as our journalists were geographically remote; both services include a screen sharing feature, allowing journalists to demonstrate aspects of their investigative process. We recorded these interviews and demonstrations using a screen capture application and later transcribed them. The deadline-driven nature of journalism precluded multiple interviews during an ongoing investigation, so we chose to interview each journalist after their investigation was complete, despite the known limitations of retrospective introspection [13]. Journalists were encouraged but not expected to keep a diary relating to their ongoing use of *Overview*. Five of our case study journalists wrote blog posts about their process, and one of them (TULSA) also sent us his personal notes.

We also collected usage logs for each journalist, consisting of timestamped interactions with *Overview*, which included selecting, viewing, and tagging documents and clusters. Log file analysis allowed us to partially reconstruct a journalist's analysis process, complementing information divulged to us in their retrospective interview. Finally, each journalist provided us with their tagged document collection, which helped to establish a shared context.

5.1 Case Studies

The six case studies we present, summarized in Table 1, took place between February 2012 and December 2013, as indicated in Figure 2.

CS1: IRAQ-SEC [50]: Our first case study took place in February 2012, when journalist and co-author Stray used *Overview v2* to analyze recently declassified documents from the Iraq war concerning the behavior of private security contractors. In particular, he wanted to categorize and count types of documented incidents involving these contractors; aside from the high-profile incidents that made headlines, he wanted to determine the prevalence of other incidents that these contractors were involved in during the Iraq war.

The document collection was the result of a FOIA request to the U.S. State Department, comprised of 666 incident reports over 4,500 pages, which were scanned using OCR. After the documents were loaded in *Overview*, Stray examined document clusters over the course of five days: he navigated the *Topic Tree*, selected clusters and their contained documents, filtered clusters using the tree pruning controls, and annotated approximately 48% of the documents with 28 unique tags. After a lengthy "orientation" phase to determine incident categories of interest, he sampled the documents using the "Select Random" button (above the *Cluster List* in Figure 4), which would select a document from the *Document List* to be shown in the *Document Viewer*. With this approach, he read and hand-coded 50 of the 666 reports, which allowed him to develop hypotheses regarding the prevalence of certain incident types. Afterward, he followed up with U.S. State Department representatives, who provided additional context and a timeline for these incidents. His published story [50] combines his categorical summarization with the context of the war.

CS2: TULSA [52]: The first case of self-initiated adoption by a journalist took place in June 2012, revealing a different motivation for using *Overview*. In this case, the journalist wanted to locate and identify evidence, documents that would support or refute a pre-existing hypothesis: he was following-up on an anonymous tip regarding municipal government mismanagement and potential conflicts of interest between city hall, municipal police, and police equipment vendors. He filed a FOIA request with the Tulsa, Oklahoma City Hall for email messages between these organizations, and then used *Overview v2* to examine 5,996 of these email messages.

His search for corroborating evidence spanned multiple sessions over 18 days, beginning with an exhaustive and systematic left-to-right navigation of the *Topic Tree*, filtering clusters using the tree pruning controls, and selecting clusters to view their contained documents. He viewed roughly 70% of the documents in the *Document Viewer* at least once, annotating 92% of them with 22 unique tags. We observed that he undertook multiple iterations of tagging: he began by tagging entire clusters using terms appearing in cluster keywords, but later tagged individual documents throughout the tree with tags such as *"important"*, *"weird"*, and *"follow-up"*. As a result of this thorough tagging, the journalist was able to lookup and browse previously identified clusters or documents of interest, focus on documents annotated by multiple tags, or locate documents that remained untagged; the latter was accomplished by selecting uncoloured points in the scatterplot. These tags also provided a starting point for the further annotation of 129 *"important"* documents with notes relating to his hypothesis; these notes eventually became integral parts of his published story [52].

CS3: RYAN [14]: In October 2012, *Overview v2* was used yet again to locate evidence in support of a hypothesis, though there are several differences as compared to the TULSA case study. In this case, a journalist wanted to follow-up on an earlier story and on accusations made by Vice President Biden that vice-presidential nominee Paul Ryan's campaign statements were hypocritical. In order to support or refute this hypothesis, the journalist sought to compare Ryan's campaign statements regarding wasteful government programs to his correspondence with various federal agencies concerning those same programs. After filing over 200 FOIA requests to these agencies, the journalist received 8,680 pages of correspondence. These physical documents arrived in several batches, and were scanned using OCR.

The journalist wanted to find genuine correspondence signed by Ryan; however, prevalent OCR errors prevented him from locating these documents using keyword search. *Overview* was able to cluster documents effectively on the remaining intact text, and most of the documents in this collection were quickly found to be irrelevant to his hypothesis. Over the course of half a day, he navigated the *Topic Tree* to locate and identify a small subset of clusters containing 176 pages of genuine correspondence containing Ryan's signature; the remainder could be safely ignored, comprised of attachments and other irrelevant correspondence. Unlike the TULSA journalist, the RYAN journalist annotated a mere 8% of the document collection with 12 unique tags. As with TULSA, the RYAN journalist used tags as a starting point for the further annotation of his source documents with notes; his published story [14] compares these findings to Ryan's campaign statements.

CS4: GUNS [29]: The first documented adoption of *Overview*'s web application deployment (*v3*) came in December 2012. Shortly after the Newtown school shooting, the journalist asked *Daily Beast* readers to self-identify as gun owners or non-owners, to report where they lived, and to post their opinion on the debate on gun ownership on a discussion board. He collected 1,278 comments: 757 from gun owners, 521 from non-owners. He aimed to determine what the debate on gun ownership is about: do gun owners and non-owners raise the same issues? He was also curious about geographical differences.

He uploaded the responses from gun owners and non-owners into two separate instances of *Overview*. Like the IRAQ-SEC case study, the GUNS journalist was interested in summarizing a document collection, though the form of this summarization was different. In IRAQ-SEC, the journalist wanted to categorize and count types of documented incidents; in contrast, the GUNS journalist sought to identify documents that were representative of their clusters, the sensational and polarizing speaking points from both sides of the debate over gun ownership; he was less interested in a fine-grained classification or quantification. For both sets of documents, he navigated and selected clusters and their contained documents, compared related clusters between the *gun owner* and *non-owner* instances, and later browsed previously identified clusters to identify representative quotes. Ultimately, he read nearly all the discussion board comments over the course of a day. Unlike the previous case studies, he did not use *Overview*'s tagging functionality, instead opting to copy quotes into an Excel spreadsheet, where he integrated geographical metadata and iteratively arranged quotes to construct a narrative for his story [29].

CS5: DALLAS: In August 2013, a journalist used *Overview v4* in a similar fashion to that of the TULSA journalist, though the outcome of their investigations differed. In the DALLAS case study, the journalist had recently reported on a collection of 4,653 email messages resulting from a FOIA request regarding a state government's response to an emergency incident. The journalist believed that some remaining evidence was left to be located, beyond what had already been reported in the earlier story. Despite having already read all the documents in the collection (unassisted by *Overview*), the journalist used *Overview* to verify that nothing was overlooked and sought to gather material for a follow-up story. She subsequently used *Overview* to examine four additional collections of messages, analyzed individually, ranging in size between 1,858 and 3,564 email messages.

The keyword search feature introduced in *Overview v4* was found to be particularly useful: the journalist alternated between identifying clusters by navigating, filtering, and selecting nodes in the *Topic Tree*, and locating documents via keyword search, then identifying related documents. As her analysis progressed, we observed that the journalist relied more upon keyword search to highlight clusters of interest within the *Topic Tree*. She applied tags to each of the five document collections: the number of tags ranged between 3 and 7, and between 7% and 52% of documents were annotated with at least one tag; in total, 14 out of 31 tags were created from keyword search results.

In this case, *Overview* was used to make the decision *not* to publish: after 12 hours of *Overview* usage spanning several weeks, the journalist was sufficiently confident that nothing significant had been overlooked in the previous investigation, ultimately deciding not to write a follow-up story. This user estimated that it would have taken "more than a week" to reach this conclusion without *Overview*, and is "definitely planning on using it again for large document sets".

CS6: NEWYORK [41]: The final case study we report took place in December 2013, in which a journalist used *Overview v4* to confirm that a document collection *did not* contain evidence that would refute his hypothesis. In the NEWYORK case study, the journalist had gathered material to investigate the state of New York's system for handling and responding to police misconduct cases, including 1,680 proposed and passed bills retrieved from the State Senate Open Legislation API. They hypothesized that the state legislature had failed to pass any bills addressing this misconduct by increasing oversight.

A considerable amount of data wrangling was required before this journalists could use *Overview*. The State Senate API provided the bills in JSON format; to address this, the journalist wrote a script to import these documents into a database, which was in turn used to export a CSV file that *Overview* could ingest.

Following data ingestion, the journalist used Overview for about 4 hours over the course of three days to read *all* the document titles and keywords in a systematic fashion: starting with the smaller nodes, he would select a node in the *Topic Tree* and scan the document titles and keywords appearing in the *Document List*; the titles tended to be verbose and descriptive, and any that were deemed interesting were read in the *Document Viewer* or tagged as *"review"* . He eventually examined the largest node, which contained 732 documents with similar titles and keywords, their contents mostly comprised of boilerplate text; the journalist tagged the entire node as *"no unless"*, meaning that any document contained by the node was not significant unless there was another tag on it. He later returned to documents tagged with *"review"*, replacing this tag with one of five descriptive tags. Though the tag highlighting used in *Overview*'s *Topic Tree* allowed the journalist to quickly locate tagged documents, he suggested that the visualization could alternatively hide all documents *not* marked with a particular tag, such as his *"not of interest"* tag.

His approach was similar to TULSA and DALLAS, in that they all sought to locate and identify clusters containing potential evidence. However, the TULSA and DALLAS journalists could have stopped their search once this evidence was found, as it is unlikely that any additional evidence would invalidate their previous findings. In contrast, the NEWYORK journalist sought to prove the *non-existence* of evidence, which required review of every document, as any evidence that went overlooked would have invalidated a claim of non-existence.

As a result of his analysis, the journalist was confident that no bills had been passed to address police misconduct, though several relevant bills had been proposed multiple times; conveniently, multiple versions of proposed bills were clustered together in *Overview*'s *Topic Tree*. While this finding is reported in a only a single paragraph of his published story [41], it played a key role in his argument that the state of New York is facing a police oversight problem; this story received considerable acclaim from the journalism community, and was a finalist for the 2014 Pulitzer Prize [4].

5.2 Think-Aloud Evaluation with Prospective Users

To complement our case study observations, we also solicited feedback from prospective journalist users. After the deployment of the web-based *Overview v3*, which included usage tracking, we observed that *Overview* and its individual features were not being used to the extent that we had hoped. We suspected usability problems so we em-

Case Study	1: IRAQ-SEC [50]	2: TULSA [52]	3: RYAN [14]	4: GUNS [29]	5: DALLAS	6: NEWYORK [41]
Date	Feb. 2012 🟣	Jun. 2012 🟢	Oct. 2012 🟢	Dec. 2012 🟢	Aug. 2013 🟢	Dec. 2013 🟢
Version	*v2* / desktop	*v2* / desktop	*v2* / desktop	*v3* / web	*v4* / web	*v4* / web
Document Collection	666 reports / 4,500 pages from FOIA (scanned using OCR).	5,996 email messages from FOIA.	8,680 pages of correspondence from multiple FOIAS (scanned using OCR).	2 collections of online discussion board comments (757 in the first, 521 in the second).	5 collections of email messages from FOIAS, ranging from 1,858 to 4,653 messages.	1,680 proposed and passed bills retrieved with NY Senate Open Legislation API.
Task	**T1**: *generate hypotheses → explore → summarize*	**T2**: *verify hypotheses → locate → identify*	**T2**: *verify hypotheses → locate → identify*	**T1**: *generate hypotheses → explore → summarize*	**T2**: *verify hypotheses → locate → identify*	**T2**: *verify hypotheses → locate → identify*
Outcome	Summarized prevalence of document categories.	Located evidence supporting hypothesis.	Located a small subset of document clusters relevant to hypothesis.	Summarized using exemplar documents.	Could not locate evidence to support hypothesis.	Proved non-existence of evidence.

Table 1. A summary of the six case studies. OCR = Optical Character Recognition; FOIA = Freedom of Information Act.; see Fig. 2 for colour coding.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2346431, IEEE Transactions on Visualization and Computer Graphics

barked on a discount usability testing program inspired by the work of Nielsen [38]: five naïve journalists were independently presented with an example document collection, such as the collection featured in Figure 4, and asked to narrate their actions as they interacted with *Overview* using a think-aloud protocol, resulting in a qualitative understanding of usability problems.

All who participated in these think-aloud sessions found *Overview* to be confusing; much of this confusion was due to the visual complexity of its multiple-view interface, as well as a lack of affordances for common and critical interactions, such as selecting a document to read. We suspect that many previous document set visualization tools would face similar usability problems in real workflows, either by lacking a robust document import feature, or by not providing a means to read individual documents [9]. An exception is *Jigsaw*, whose developers have noted and overcome similar problems [17]. In the next section, we discuss how the design of *v4* resolved these usability problems.

6 Analysis

Given our observations or real world usage, we now revisit our initial task abstraction and discuss the rationale for *Overview*'s design.

6.1 Task Abstractions Reconsidered

After the GUNS case study, we struggled to distinguish between journalists' goals, approaches, and outcomes. Specifically, the TULSA and RYAN journalists used *Overview* in more directed and systematic ways that we did not anticipate, in that they sought to locate specific evidence or a subset of clusters that were relevant to a pre-existing hypothesis, forcing us to reconsider our initial task abstraction of *"exploring"* a document cluster structure, which was based on the WARLOGS use case. A number of previous tools aim to help the user *"explore"* a document collection (such as [6, 9, 10, 12]), though few of these tools have been evaluated with users from a specific target domain who bring their own data, making us suspect that this imprecise term often masks a lack of understanding of actual user tasks.

In 2013, we developed and proposed a typology of abstract visualization tasks [5], the purpose of which was precisely to articulate such differences in visualization usage at multiple levels of specificity. According to this typology, a task description is broken down into *why* a task is undertaken, *what* dependencies a task might have, and *how* the task is supported. *How* is somewhat orthogonal to *why*, as exemplified by the differences in usage reported in the previous section. We applied this typology to the coding of our observational data, identifying two different tasks, **T1** and **T2**, that replace and improve upon our initial task abstraction. In this section, we will use the vocabulary and notation of this typology to focus on *why* and *what*; in Section 6.2, we analyze *how Overview* supports these tasks.

T1: generate hypotheses → explore → summarize: When approaching a collection of leaked documents or a corpus of social media content, a journalist may have little prior knowledge regarding the collection's content, eliciting a need to *generate hypotheses* and to ask *"what's in this collection?"*. To support the generation of hypotheses, a journalist must be able to *explore* a document collection and *summarize* clusters of documents. The term *explore* is defined more precisely in our typology as a form of *search* in which neither the identity nor the location of a search target are known a priori. In the context of a document collection, a search target may be content within a document, an individual document itself, a cluster of related documents, or an arbitrary set of documents and clusters. *Exploring* is distinguished with *browsing*, in which the location of a search target is known but the identity is not, *locating*, in which the converse is true, and *lookup*, in which both location and identify are known. The result of *summarizing* is a compressed representation of the full contents of the document collection, such as the categories and counts produced in the IRAQ-SEC case study, or the exemplar documents that the journalist ultimately selected in GUNS case study.

T2: verify hypotheses → locate → identify: In contrast, a journalist who asks for documents via FOIA request typically has some pre-existing hypotheses, and their aim is to *verify*, *refute*, or *refine* these hypotheses by *locating* evidence. In these cases, a journalist likely has a sense of what the documents are about, but they may not be able to specify the evidence they seek in terms of a standard search query (for example, "corruption" would not suffice), and there may also be unexpected but valuable material waiting to be discovered. In the language of our typology, the aim is to *locate* and *identify* clusters containing potential evidence, beginning with those labelled by interesting keyword terms; alternatively, a journalist will *locate* documents containing specific search terms and subsequently *browse* and *identify* related documents found in the same cluster. **T2** describes the use of *Overview* in the TULSA, RYAN, DALLAS, and NEWYORK case studies.

Throughout both **T1** and **T2**, a journalist will often *produce* notes or annotations for documents as they *generate*, *verify*, or *refine* their hypotheses, perhaps *comparing* documents to each other or to secondary sources outside the collection.

6.2 Design Rationale

With a more precise understanding of journalists' tasks, we now analyze the rationale for our visual encoding and interaction design choices, with the intent of transferability to other domain problems involving similar data and task abstractions [46].

Why show a tree? *Trees afford structured and systematic exploration.* When a document collection contains separable clusters of similar documents, a tree visualization affords a systematic and, if desired, exhaustive traversal of these clusters. The TULSA case study is an example where the journalist based his choice of which documents to read based on the structure of the tree, sweeping from left to right. The *Topic Tree* also includes a visual encoding of applied tags, which makes it possible for the user to identify the documents they have and have not already tagged, and how tags correspond to clusters.

How to show a tree? *Emphasize interior nodes (not edges or leaves); instil trust in the underlying algorithm.* The *Topic Tree* in the first two versions of *Overview* (Figure 4) rendered all clusters as identical nodes. While tree visualizations are often associated with the task of path tracing and determining connectivity [33], *Overview* users are primarily interested in the properties of nodes corresponding to document clusters, such as the number of documents contained by a cluster, or the key terms that describe these documents.

The *Topic Tree* of *v3-4* directly encodes cluster size as node width; this design choice allows the user to compare cluster sizes directly, or work systematically from larger to smaller topics, of particular use when trying to summarize a document collection to some desired degree of detail (**T1**). In enlarging the width of nodes, it also became possible to encode the number of documents tagged within the cluster as a colour label having a width proportional to the size of the node, as shown in Figure 5. We opted not to use a space-filling treemap visualization of hierarchical document clusters because this approach would place too much emphasis on leaf nodes; when summarizing a collection (**T1**) or when locating a subset of documents (**T2**), the mid-level interior nodes in the tree are typically the most informative. While less space efficient, a tree with variable-width nodes provides more flexibility, especially given the differences between **T1** and **T2**.

With larger nodes, we were able to display cluster keyword terms directly in the node itself, rather than in a separate *Cluster List* view, as in *v1-2*, which displayed keywords only for the selected cluster and its descendants. Displaying keywords within nodes allows users to compare keywords at a glance, both between and within clusters.

The design of the *Topic Tree* required a balance between usability and cluster fidelity: in *v3*, there was no limit on the number of children allowed for each node, a situation reported to be overwhelming by journalists who participated in the think-aloud evaluation. When a node can have many children, tree exploration reduces to linear search; for this reason, the maximum number of child nodes was limited in *v4* by switching to a recursive adaptive *K*-means algorithm [42] with an upper limit of five children per node.

We added explicit cluster fidelity labels to the *Topic Tree* nodes in *v4* to help users interpret the content of a cluster: the labels *"Some"*, *"Most"*, and *"All"* show how many documents in a cluster contain each keyword and thus signal the consistency of topics found in that node, as shown in Figure 5. These labels help the user to decide whether to treat the node as a conceptual unit that might be tagged as a whole, or expand it to examine its children individually. They help the user assess cluster consistency and separability and serve to build trust in the clustering algorithm [8]. Previously, users had to judge the topical consistency of a cluster by examining the individual documents within it, or by referring to the scatterplot in a way that is known to be difficult for dimensionally-reduced data [47].

How to interact with a tree? *Selective pruning and informative tooltips.* In *v1-v2*, the user was able to clarify the *Topic Tree* by pruning (filtering) small nodes, according to a threshold selected from a set of seven coloured radio buttons below the *Topic Tree*. Our case studies revealed that many users never understood that the variable tree-pruning threshold used in *v1-v2* was hiding nodes from them, a problem especially for those intent on locating evidence or proving the non-existence of evidence (**T2**). We replaced threshold-based node pruning with a selective expand/collapse option on each node; when combined with panning and zooming, these interactions provide users with a fine-grained control over focus and context.

Upon selecting a node in *v1-v2*, keywords for the selected cluster and its descendants were shown in a status bar between the *Topic Tree* and the *Cluster List*, however these were spatially removed from the user's point of focus. To resolve this, we added tooltips in *v3* that show cluster keywords when the cursor hovers over a node.

Why no scatterplot? *Unstructured exploration is redundant for* **T1** *and* **T2**. Using scatterplots to visualize document collections is an approach common to previous work [7, 18, 12, 2, 40]. We thought that a scatterplot would allow users to judge cluster size, quality, and separability [47]. However, scatterplots do not directly show cluster content, such as document keywords, unless tooltips or point aggregation is used. We did not pursue the use of these techniques because we discovered that the scatterplot was seldom used in the case studies of journalists who adopted *Overview*. The TULSA journalist was an exception in that he used the scatterplot to locate untagged documents containing potential evidence after extensive use of the *Topic Tree* (**T2**). This task would have been better served by providing a direct way to show how many untagged documents a node contains; the *"Show Untagged"* button introduced in *v4* accomplishes this.

Ultimately, we realized that a scatterplot does not help users overcome the burden of choice overabundance when determining which cluster to investigate next [45], whereas the tree-based hierarchical clustering used in the *Topic Tree* affords a form of structured navigation. In addition, the cluster fidelity labels introduced in *v4* help users to assess cluster consistency and separability, thereby eliminating any further need for a scatterplot.

Why tags? *Tags provide simple annotation, progress tracking, and user-defined semantics.* Tagging was used extensively in five of the six case studies. Some tags aligned with cluster boundaries (IRAQ-SEC, RYAN, and the first set of tags created in TULSA), while other tags appeared throughout the tree (DALLAS, NEWYORK, and the second set of tags created in TULSA). Tags are a simple and flexible form of annotation that help users track where they have been and what they have learned. They can also be used to impose a context-specific organization scheme on a document collection. No single clustering will meet all analysis needs, since any high-dimensional dataset is likely to have multiple cross-cutting semantically interesting clusterings [19]. The "best" clustering will depend on the documents and the story. *Overview* does not support manual re-arrangement of the *TopicTree* hierarchy, as in systems such as *HierarchicalTopics* [10]. Instead, we support manual tagging as a simple and flexible way for the users to impose their own semantics on a document collection, where the cluster structure can be leveraged as a useful scaffold when it matches user semantics, but ignored when it does not. The most recent feature added to *Overview*, developed after the case studies in this paper, supports the creation of multiple trees, giving different views of same document collection. The user can control the clustering by entering words to ignore, which prevents *Overview* from clustering based on document letterhead or boilerplate text, and by entering especially important words which are weighted higher when constructing document vectors. It is also possible to create a tree containing only a subset of the documents, specified by selecting an existing tag.

Multiple views: how many and how to coordinate? *Less is more, provide obvious affordances.* The evolving design of *Overview* recalls the challenges and considerations for designing multiple view systems [32]. These considerations include *how many discrete views are appropriate?*, *how should views be arranged?* and *how should views be coordinated?*. A consensus on these questions has not yet been reached, as multiple-view visualization systems range from dual-view to over 20, with a similar range in view coordination patterns [53].

The CARACAS pilot case study with *Overview v1* revealed that the *Document Viewer* was too small and the selection of documents and clusters across the different views was poorly coordinated. Despite improvements to view coordination in *v2* and *v3*, the views were not always well understood; for example, those who participated in usability testing did not initially realize that the *Document List*, displayed as a line of keywords for each document, was in fact a list of selectable documents. In *v4*, *Overview*'s interface was streamlined into three views coordinated with linked selection and highlighting: the *Topic Tree*, a consolidated *Document Viewer/List* featuring document titles and list navigation controls, and a list of *Tags*; as described above, we removed the scatterplot visualization and *Cluster List*, both made redundant by the redesigned *Topic Tree*. While we might have made *Overview* a single-view system, we instead reasoned that having the *Topic Tree* visible provides helpful context when reading a document and deciding what to read next.

How to support user workflow? *Simplify for infrequent use, reduce data wrangling.* Our findings show that simplicity and learnability are critical for journalists, because any one journalist only deals with large document collections intermittently.

After *Overview v2* was deployed and promoted within the journalism community, it became clear that many prospective users had great difficulty downloading, installing, and configuring it. Additionally, *v2* could only import documents in a CSV file; we quickly learned that journalists receive document collections in every conceivable format, from stacks of paper to database dumps. We confirmed that the need to wrangle data into compatible formats is a considerable barrier to adopting a visualization tool into an analysis workflow, as discussed in a recent research agenda [27]. We should not expect journalists to write custom data wrangling scripts, as the NEWYORK journalist had to do. To minimize the amount of configuration and wrangling required, the web-based *Overview v3-v4* required no user installation and supported import from a folder of PDF documents or from DocumentCloud [1], a document hosting service used by journalists which can itself ingest a wide variety of formats. Without this integration, we suspect that the DALLAS journalist would have been unable to make use of *Overview*. The DocumentCloud interface is integrated into *Overview*'s *Document Viewer*, which includes a function for annotating documents with notes.

We also added full-text keyword search in *v4*, as prospective users and our case study journalists had expressed a desire to flexibly alternate between locating clusters in the *Topic Tree* and a directed search for locating documents of interest, without having to use a search tool such as grep or DocumentCloud's search interface. We expect that the TULSA and RYAN journalists, both performing **T2** with *v2*, would have benefited from this keyword search feature, and that the absence of this feature would have been a deterrent for the DALLAS and NEWYORK journalists, who also performed **T2**.

We note that the use of *Overview* forms part of a larger investigation and reporting workflow: each case study journalist combined its use with many computer-assisted and non-computer-assisted methods for data collection, data transformation, and eventual story presentation.

7 Discussion

In this section, we discuss the value and logistics of conducting design studies involving multiple deployments and the analysis of user

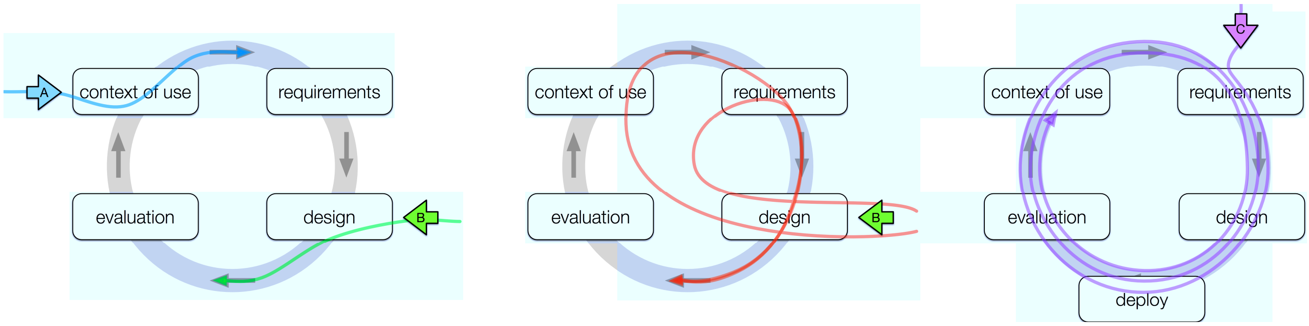


Fig. 6. The human-centred design process development cycle, in which Lloyd and Dykes [36] discern between alternative entry points (A,B) and between traditional (green), grounded (blue), and their own approach in which example designs are used to establish context of use and elicit requirements (red). In contrast, we begin with some requirements at point C, and only after multiple deployments do we arrive at a clear understanding of context of use (purple). Figure adapted and extended from [36].

adoption, as well as the limitations of this type of research.

Why study adoption? As with any iterative human-centred design process, it is difficult to know when to declare success; we consider adoption defined as repeated instances of self-initiated use to be a form of success. Adoption is particularly interesting in the domain of journalism because tool use is a decision made separately by each journalist for each story, rather than dictated by a central authority.

Though the design study process is cyclical and may include multiple deployments [46], there are surprisingly few papers that comment on the adoption of a visualization tool without the prompting of designers: in a recent survey of eight hundred visualization papers containing an evaluation component, only five commented on adoption [31]. Of these, two provide a thorough description of who adopted their visualization tool, how it had been used, whether it was still in use, and what problems users reported [30, 37]. Our work adds to this short list, as does the recent study of *Jigsaw*'s adoption [28].

Many design studies report on deployment to a target group of users and evaluate their reaction to the tools during a period of intense study, but our own experience and personal communication with other practitioners leads us to believe that visualization tool use typically drops off after a paper is submitted. Gonzales and Kobsa provide a rare example of explicitly checking back after this time period; they found that their target users did not adopt the visualization system despite the promising initial results reported in their original paper, and conjecture that a misunderstanding of user workflow was the primary factor behind this lack of adoption [15]. We conjecture that this situation might be the common case, and thus that longer-term adoption rates may be very low for research prototypes. Sustained follow-up by researchers until adoption is achieved provides a way to disambiguate whether the barrier to adoption is truly only a workflow issue, or an indication that the tool failed to address the true needs of the target users.

The logistics of studying adoption: Before *Overview* was deployed, we could not have fully predicted *why* and *how* journalists would approach large document collections; we were unable to verify the correctness of our task and data abstractions. *Overview* is sufficiently novel that its value could not be assessed without adequately functional prototypes. We argue that this situation is common in visualization because of the complexity of the data and tasks at play, recalling the argument of Lloyd and Dykes about the need for *data sketches* [36]: functional example designs for establishing context of use and eliciting requirements. They contrasted their design-first approach, illustrated by the red trajectories in Figure 6, to the traditional design-then-evaluate approach (the green trajectory) and to an approach grounded in user context [26] (the blue trajectory). The purple trajectory illustrates that close collaboration with domain experts from the very start of a project means they bring expertise about requirements to the table, so it is not a design-first endeavour. The multiple loops in the purple trajectory emphasize the importance of deploying a visualization tool as a precursor to evaluation, and we note that after these loops the trajectory ends at context of use: it took several deployments and case studies of self-initiated journalists who adopted *Overview* before we attained a clear understanding of users' tasks and their broader analysis workflows.

Our use of case studies to study adoption is methodologically similar to qualitative longitudinal evaluation studies described in previous work [36, 44, 48]. Our approach differs from these in that we engaged a different set of users at each stage of design, rather than the same set of users. This difference reflects *Overview*'s context of use: repeat usage cannot be predicted and *Overview* is only appropriate for some investigations; we have yet to encounter a journalist who specializes in investigations pertaining to large document collections.

Limitations and future work: A limitation of adoption-phase research is that a set of specific target users cannot be identified in advance, in contrast to the typical design study chronology [46]. As a result, there is an inherent selection bias in our case studies, because they largely represent successful cases; a similar observation was made by McKeon, who interviewed only prolific users of his deployed visualization tool [37]. In future work, we would like to know more about cases in which *Overview* was used briefly and then abandoned as being unsuitable for the problem at hand.

To broaden our understanding of how *Overview* is used, we hope to investigate the use of *Overview* in other domains where large document collections are prevalent, such as intelligence analysis [28], law [20], medicine, and digital humanities research; our set of task abstractions may continue to expand. Meanwhile, we are continuing to monitor and learn from new cases of adoption by journalists.

8 Conclusion

We presented a visualization design study of *Overview*, an application for the systematic analysis of large document collections. *Overview* has proven to be useful, having been used in investigations leading to at least nine published stories [3]. Using a recently proposed typology of visualization tasks [5], we identified two task abstractions based on findings from six case studies. Given our data and task abstractions, we rigorously analyzed the effectiveness of *Overview*'s visual encoding and interaction design. This analysis generalizes beyond the domain of journalism, and speaks to the design and evaluation of other visualization tools for supporting the analysis of document collections and clustered dimensionally-reduced data in general. Finally, this work adds to the small number of studies found in the visualization literature that include observations of user adoption; in observing real world usage by self-initiated journalists, we confirmed that several iterations of design and deployment are required before fully understanding *why* and *how* a visualization system will be used in practice.

Acknowledgments

We received financial support from the Knight News Challenge and from NSERC. We thank J. Karlsson and A. Hooper, who contributed to the development *Overview v3-v4*. We thank the case study journalists, including the DALLAS journalist from the Dallas Morning News who requested to remain anonymous. We also thank J. Dawson, J. Ferstay, H. Lam, J. McGrenere, R. Rensink, and M. Sedlmair for their comments on the project and paper.

References

[1] DocumentCloud. http://documentcloud.org/.

[2] Leaksplorer Beta. http://leaksplorer.org/.

[3] Overview: Completed Stories. http://overview.ap.org/completed-stories/.

[4] The Pulizer Prizes: 2014 Finalists. http://pulitzer.org/finalists/.

[5] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE TVCG (Proc. InfoVis)*, 19(12):2376–2385, 2013.

[6] A. J. B. Chaney and D. M. Blei. Visualizing topic models. In *Proc. Intl. AAAI Conf. Weblogs and Social Media (ICWSM)*, pages 419–422, 2012.

[7] Y. Chen, L. Wang, M. Dong, and J. Hua. Exemplar-based visualization of large document corpus. *IEEE TVCG (Proc. InfoVis)*, 15(6):1161–1168, 2009.

[8] J. Chuang, D. Ramage, C. D. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proc. ACM Conf. CHI*, pages 443–452, 2012.

[9] W. Cui, S. Liu, L. Tan, S. C, Y. Song, Z. J. Gao, X. Tong, and H. Qu. TextFlow: Towards better understanding of evolving topics in text. *IEEE TVCG (Proc. InfoVis)*, 17(12):2412–2421, 2011.

[10] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE TVCG (Proc. VAST)*, 19(12):2002–2011, 2013.

[11] J. Eisenstein, D. H. P. Chau, A. Kittur, and E. P. Xing. TopicViz: Interactive topic exploration in document collections. In *Proc. Extended Abstracts ACM Conf. CHI*, pages 2177–2182, 2012.

[12] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE TVCG (Proc. VAST)*, 18(12):2879–2888, 2012.

[13] K. A. Ericsson and H. A. Simon. Verbal reports as data. *Psychological Review*, 87(3):215–251, 1980.

[14] J. Gillum. Ryan asked for federal help as he championed cuts. *Associated Press*, Oct. 12, 2012. http://goo.gl/RM7uk.

[15] V. M. Gonzalez and A. Kobsa. A workplace study of the adoption of information visualization systems. In *Proc. Intl. Conf. Knowledge Management*, pages 92–102, 2003.

[16] C. Görg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. T. Stasko. Combining computational analyses and interactive visualization for document exploration and sensemaking in Jigsaw. *IEEE TVCG*, 19(10):1646–1663, 2013.

[17] C. Görg, Z. Liu, and J. Stasko. Reflections on the evolution of the Jigsaw visual analytics system. *Information Visualization*, (in press).

[18] M. Granitzer, W. Kienreich, V. Sabol, K. Andrews, and W. Klieber. Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In *Proc. IEEE Symp. InfoVis*, pages 127–134, 2004.

[19] J. Grimmer and G. King. General purpose computer-assisted clustering and conceptualization. *Proc. National Academy of Sciences (PNAS)*, 108(7):2643–2650, 2011.

[20] M. R. Grossman and G. V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, XVII(3):1–33, 2011.

[21] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing thematic changes in large document collections. *IEEE TVCG*, 8(1):9–20, 2002.

[22] M. A. Hearst and D. Rosner. Tag clouds: Data analysis tool or social signaller? In *Proc. Hawaii Intl. Conf. System Sciences (HICSS)*, 2008.

[23] E. Hetzler and A. Turner. Analysis experiences using information visualization. *IEEE CG&A*, 24(5):22–26, 2004.

[24] S. Ingram. *Practical Guidance for Dimensionality Reduction: User Guidance, Costly Distances, and Document Data*. PhD dissertation, University of British Columbia, 2013.

[25] S. Ingram, J. Stray, and T. Munzner. Hierarchical clustering and tagging of mostly disconnected data. Technical report, University of British Columbia, 2012. http://goo.gl/Il0pUU.

[26] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale. Grounded evaluation of information visualizations. In *Proc. ACM BELIV Workshop*, 2008.

[27] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[28] Y. A. Kang and J. T. Stasko. Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts. *IEEE TVCG (Proc. VAST)*, 18(12):2869–2878, 2012.

[29] M. Keller. Own a gun? Tell us why. *The Daily Beast*, Dec. 22, 2012. http://goo.gl/2yRgk.

[30] R. Kincaid, A. Ben-Dor, and Z. Yakhini. Exploratory visualization of array-based comparative genomic hybridization. *Information Visualization*, 4(3):176–190, 2005.

[31] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE TVCG*, 18(9):1520–1536, 2012.

[32] H. Lam and T. Munzner. A guide to visual multi-level interface design from synthesis of empirical study evidence. *Synthesis Lectures on Visualization*, pages 1–117, 2010.

[33] B. Lee, C. Plaisant, C. S. Parr, J. D. Fekete, and N. Henry. Task taxonomy for graph visualization. In *Proc. ACM BELIV Workshop*, 2006.

[34] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. TIARA: Interactive, topic-based visual text summarization and analysis. *ACM Trans. Intelligent Systems and Technology*, 3(2), 2012.

[35] Y. Liu, S. Barlowe, Y. Feng, J. Yang, and M. Jiang. Evaluating exploratory visualization systems: A user study on how clustering-based visualization systems support information seeking from large document collections. *Information Visualization*, 12(1):25–43, 2013.

[36] D. Lloyd and J. Dykes. Human-centered approaches in geovisualization design: investigating multiple methods through a long-term case study. *IEEE TVCG (Proc. InfoVis)*, 17(12):2498–2507, 2011.

[37] M. McKeon. Harnessing the web information ecosystem with wiki-based visualization dashboards. *IEEE TVCG (Proc. InfoVis)*, 15(6):1081–1088, 2009.

[38] J. Nielsen. Why you only need to test with 5 users, March 19, 2000. http://goo.gl/6Huppn.

[39] P. Österling, G. Scheuermann, S. Teresniak, G. Heyer, S. Koch, T. Ertl, and G. H. Weber. Two-stage framework for a topology-based projection and visualization of classified document collections. In *Proc. IEEE Symp. VAST*, pages 91–98, 2011.

[40] F. V. Paulovich, M. C. F. Oliveira, and R. Minghim. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Proc. IEEE Brazilian Symp. Computer Graphics and Image Processing*, pages 27–36, 2007.

[41] S. Peddie and A. Playford. For their eyes only: Police misconduct hidden from public by secrecy law, weak oversight. *Newsday*, Dec. 28, 2013. http://goo.gl/KBT7CA.

[42] D. Pham, S. Dimov, and C. Nguyen. Selection of K in K-means clustering. *J. Mechanical Engineering Science (Proc. IMechE)*, 219:103–119, 2005.

[43] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[44] P. Saraiya, C. North, V. Lam, and K. Duca. An insight-based longitudinal study of visual analytics. *IEEE TVCG (Proc. InfoVis)*, 12(6):1511–1522, 2006.

[45] B. Schwartz. *The Paradox of Choice: Why More Is Less*. Ecco, 2003.

[46] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE TVCG (Proc. InfoVis)*, 18(12):2431–2440, 2012.

[47] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE TVCG (Proc. InfoVis)*, 19(12):2634–2643, 2013.

[48] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proc. ACM BELIV Workshop*, 2006.

[49] J. Stray. A full-text visualization of the Iraq War Logs. *Associated Press*, Dec. 10, 2010. http://goo.gl/IITFpE.

[50] J. Stray. What did private security contractors do in Iraq? *Associated Press*, Feb. 21, 2012. http://goo.gl/qqGGh.

[51] M. Tory, C. Swindells, and R. Dreezer. Comparing dot and landscape spatializations for visual memory differences. *IEEE TVCG*, 15(6):1033–1039, 2009.

[52] J. Wade. TPD working through flawed mobile system. *Tulsa World*, Jun. 3, 2012. http://goo.gl/yE98e2.

[53] C. Weaver. Patterns of coordination in Improvise visualizations. In *Proc. SPIE–IS&T Visualization and Data Analysis (VDA)*, 2007.