Humanities Research Software Design: The Wilde Trials Web App

Scholarly and Research Communication
VOLUME 7 / ISSUE 2 / 2016

Colette Colligan, Michael Joyce, & Cécile Loyen
Simon Fraser University

Sarah Bull
Cambridge University

Oscar Wilde
Photo: Napoleon Sarony (1882)

Abstract

Background: This article discusses the design of Web-based research software to computationally analyze the international news coverage of the playwright Oscar Wilde's 1895 sex trials. Over two months, Wilde stood three trials, eventually being convicted of "gross indecency" (1885 Criminal Law Amendment Act).

Analysis: Over the past year, we have collaboratively designed a program to advance our understanding of the trials' cultural impact as they were reported in newspapers around the world. Bridging our expertise in nineteenth-century cultural history and software engineering, we discuss the concept and design of the Wilde Trials Web App, as well as early discoveries about the French news coverage and plans for the program's further development.

Conclusion and implications: Our work stands at the forefront of software design and data-driven research on the nineteenth-century press.

Keywords: Oscar Wilde; Gross indecency; Trials 1895; International; Press coverage; News; Reprint; Verbatim; Web app; Transcription; Digitize; Collection; Algorithm

Colette Colligan is Professor of English, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6. Email: ccolliga@sfu.ca .

Michael Joyce is Web and Data Service Developer in the Digital Humanities Innovation Lab, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6. Email: mjoyce@sfu.ca .

Sarah Bull is a Wellcome Trust Research Fellow in the Department of History and Philosophy of Science, Free School Lane, Cambridge, UK CB2 3RH. Email: sb2103@cam.ac.uk .

CISP Press
Scholarly and Research Communication
Volume 7, Issue 2, Article ID 0201256, 15 pages
Journal URL: www.src-online.ca
Received June 1, 2016, Accepted September 15, 2016, Published November 8, 2016

Colligan, Colette, Joyce, Michael, Bull, Sarah, & Loyen, Cécile. (2016). Humanities Research Software Design: The Wilde Trials Web App. *Scholarly and Research Communication, 7*(2): 0201256, 15 pp.

Cécile Loyen is a research assistant in the Department of English, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6. Email: cecileloyen@gmail.com .

The origins of the Web app

The Wilde Trials Web App was designed in response to a specific research question related to the famous sex trials of Oscar Wilde: to what extent were news reports on the trials reprinted per verbatim? His three trials were an international media event in 1895 and were front-page news the world over for nearly a two-month period. His celebrity and the sexual nature of the trials (which disclosed sodomy, male prostitution, sexual blackmail, intergenerational relationships, mock-marriages, cross-dressing, sexual seduction, and "the love that dare not speak its name") triggered this media sensation as well as debates about press standards. Although the Wilde trials are perhaps the most studied media event of the nineteenth century, the volume of their worldwide press coverage remains unknown (Cohen, 1993; Erber, 1996; Foldy, 1997; Fotheringham, 2003; Ivory, 2012; Robinson, 2015; Walshe, 2005; Wan, 2006). Yet, it is only when the news reports are aggregated that one begins to realize how much of the reporting within regional zones was exactly the same and circulated within press networks. The Parisian press coverage of the trials is just one example of this homogenization of news. Eighteen different daily newspapers covered the trials, publishing a total of 466 reports and over 200,000 words. Almost half of those reports printed content that was over 60 percent identical to content published by one or more competing newspapers. This is a striking discovery, one that counters assumptions about the virtuosity of the Parisian press, and opens up inquiry about the public impact of multiple exposures of the same news content from a closely connected group of newspapers. This line of inquiry can only be undertaken, however, with the collection and transcription of news reports in aggregate, and a computational instrument that can detect matching news.

The digitization of newspaper collections, by commercial providers and national libraries, has advanced capabilities for the aggregation of full-text corpuses. The bigger challenge is finding a text-sharing tool that can operate on a large corpus. There are Web-based text-comparison tools that assist with the close analysis of textual matching and difference. **Mergely** enables side-by-side comparison of two documents and visualizes textual deletions and additions. **Juxta** is an open source tool created by the Applied Research and Patacriticism team at the University of Virginia to aid the scholarly collation of digitized texts. It allows for the comparison of two or more witnesses and displays differences through three available visualizations. **CollateX**, developed as an Interedition project, also lets users compare textual variations and outputs results through variant graphs (Dekker, van Hulle, Middell, Neyt, & van Zundert, 2014; Ross & Sayers, 2014). These tools were designed for the close analysis of different versions of text, not for the large-scale multidirectional comparison of thousands of documents. Plagiarism-detection software such as Turnitin, which is licensed via subscription to educational providers, is more suitable for identifying text sharing in a large corpus. Turnitin indexes a large number of documents, detects and quantifies the percentage of text sharing across these documents, and also provides a visual interface for close textual comparison (Shahabi, 2012). Because this software is designed primarily to detect instances of plagiarism, however, its interface for close text comparison is not designed for careful and sustained analysis. A pop-up window is generated for a document in question that highlights any passages of matched text, but it requires the user to navigate between links and windows to cross-reference matching text. There is also the problem of using software for purposes other than those originally intended. Every document uploaded into Turnitin is compared to those in the account holder's corpus, the company's archive, and on the Web – expanding the scope of comparison beyond a self-contained corpus. The company also stores in perpetuity all documents uploaded on its server, which prevents the reindexing of the corpus as the software will generate match reports between the account holder's original and revised corpus, becoming a relentless self-duplicating machine.

In short, there was no existing computational instrument that could help answer the question as to what extent news reports on the Wilde trials were reprinted per verbatim. The existing tools, however, provided models for designing research software that could analyze a corpus of news reports on the Wilde trials. What was needed was a program that could house an extensible database of news reports, detect and quantify the degree of text sharing across these documents, and provide a visual interface for the close examination of shared passages. The creation of this program would enable data-driven inquiry of Wilde news reports as informational strings structured by the period's international news markets (Barth, 2014; Silberstein-Loeb, 2014) and allow for text-driven inquiry, central to traditional textual criticism, cognizant of the unique local properties of all texts and the myriad social codes embedded in their textual conditions and transformations (McGann, 1991). Such a program would ideally allow for both quantitative and qualitative analysis of these news reports – large-scale analysis of aggregated news and its potential impacts, and close analysis of textual difference, such as divergent news as well as instances of censorship and contra-censorship. The next step was to custom design such a program.

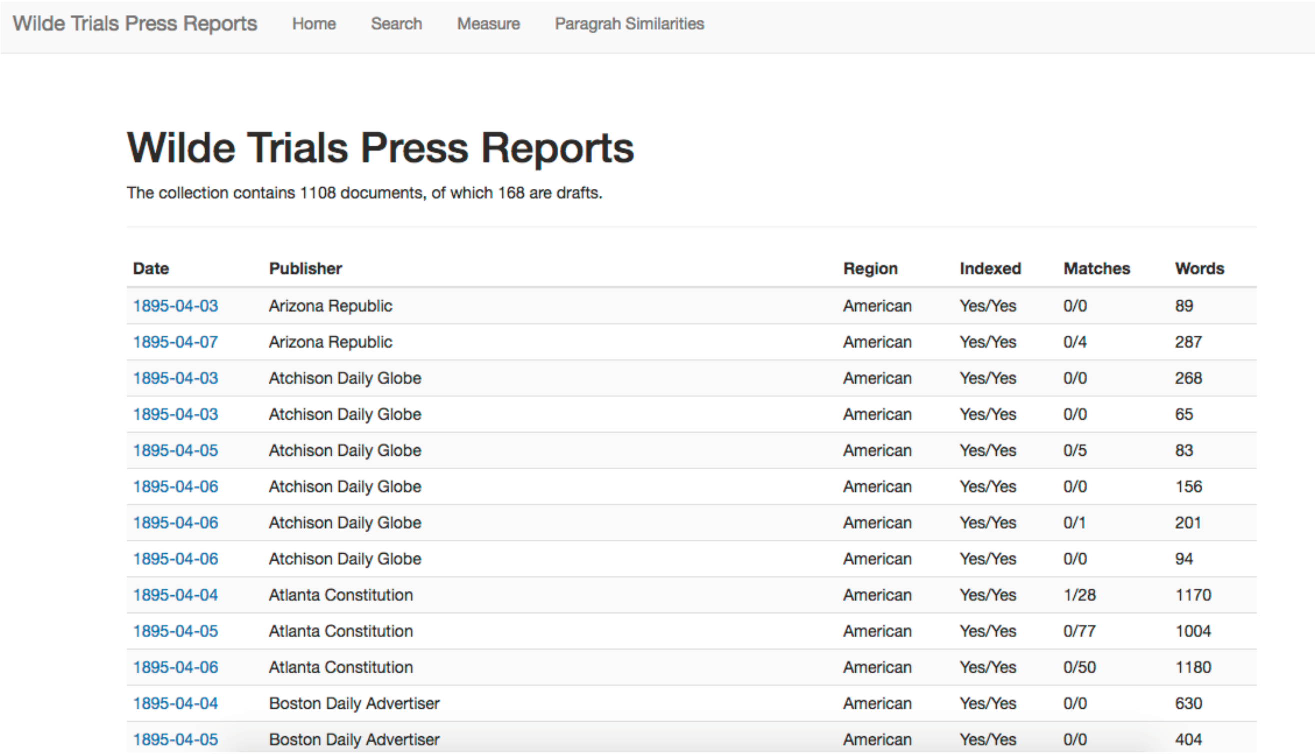
Designing the Web app

Simon Fraser University's Digital Humanities Innovation Lab, housed in the library, has been the digital laboratory for the development and design of this program. Over the last two years, it has provided optical character recognition (OCR) digitization services, technical expertise and systems resources, and opportunities for interdisciplinary collaboration. This collaboration has resulted in the Wilde Trials Web App, a custom-built textbase and text-sharing detection program for analyzing the Wilde trials news reports.

The Wilde Trials Web App currently contains over 1,100 separate news reports on the Wilde trials that have been gathered from British, French, American, and Australian newspapers, in both English and French languages. These reports are full-runs of newspaper coverage on the Wilde trials through April and May 1895; and they are all in text format and have been (or are in the process of being) double corrected by a research team. National digitization projects, such as Trove, Europeana, and Gallica, have been instrumental in this curation process, as have newspapers digitized by Gale-Cengage and other commercial providers. The program also includes transcribed reports from newspapers that have yet to be digitized, such as the expatriate English-language Parisian papers. The research team is a small and changing group of postgraduate, graduate, and undergraduate students that has worked together to locate, correct, and sometimes transcribe these reports. At this point, Dr. Sarah Bull and Cécile Loyen have been most involved in this curatorial and editorial work, the first focusing on a sample set of British reports, and the latter focusing on French reports. This corpus is continually growing: its size is limited only by the extent of newspaper digitization projects, the willingness of the bleary-eyed research team to correct (and retranscribe) OCR documents, and the scope and staging of the work. So far the group has concentrated on collecting news reports published in France and America, the two regions outside Britain where Wilde was most famous. As the only full-text report textbase – in print or online – our Web application is instrumental to large-scale research on the international coverage of the Wilde trials.

The program also includes a text-sharing algorithm that processes all of these news reports: it parses up the documents and generates data on the number of matching documents and the percentage of text sharing. Designing the Web app's text-sharing algorithm was one of the principal challenges. The main objective was to detect text sharing across multiple documents in many-to-many relationships. One news report could share a substantial amount of copy with eight different newspapers, each of which might have made subtle textual changes in the transmission and editorial process that reveal different news sources or news alliances. The program needed to capture all of those instances of matching. Choosing the right string metric to detect this text sharing happened early in the program's development stage, and was far less complicated than determining how to markup the text to maximize the algorithm's efficiency. These nineteenth-century newspapers often had shared content, but arranged it differently, depending on national practices and the material properties of the newspaper (such as paper size). Some newspapers broke up large blocks of text into smaller paragraphs; others added a series of crossheads; and others reorganized the order of paragraphs. Although the paragraph appears to have been the principal

Figure 1: Wilde Trials Web App index view



Wilde Trials Press Reports

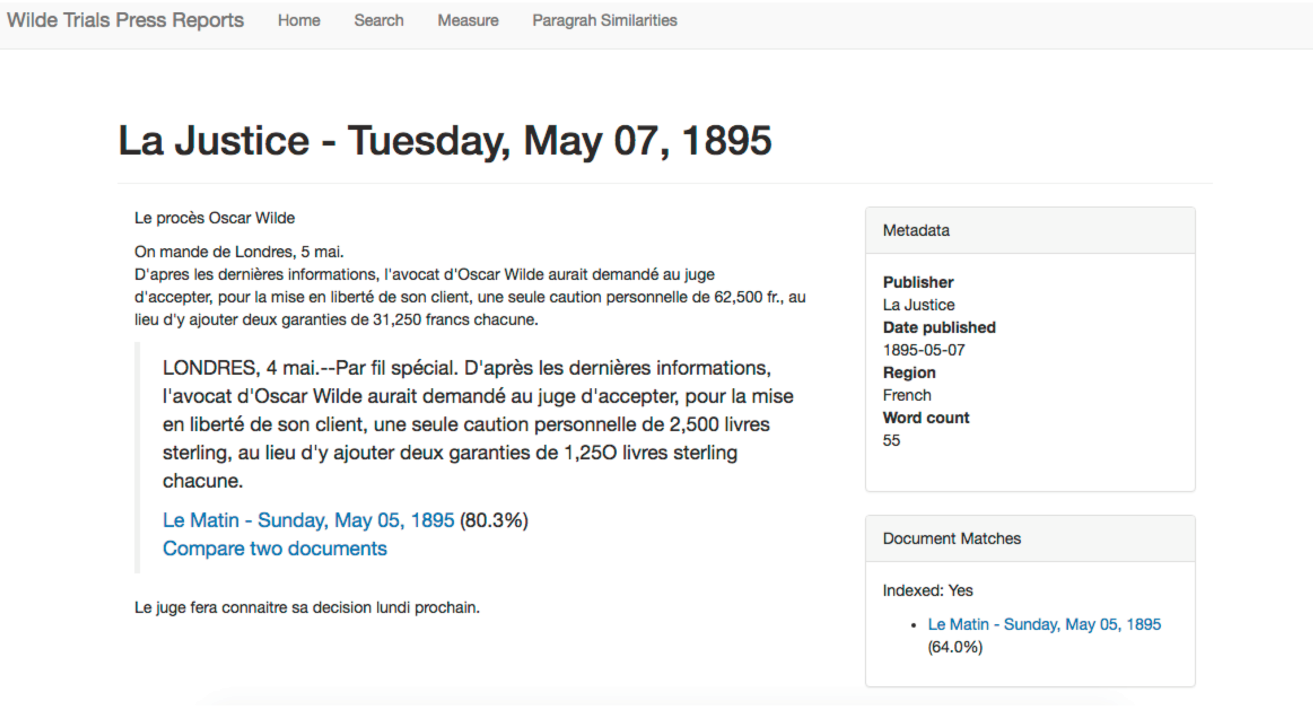
The collection contains 1108 documents, of which 168 are drafts.

Date	Publisher	Region	Indexed	Matches	Words
1895-04-03	Arizona Republic	American	Yes/Yes	0/0	89
1895-04-07	Arizona Republic	American	Yes/Yes	0/4	287
1895-04-03	Atchison Daily Globe	American	Yes/Yes	0/0	268
1895-04-03	Atchison Daily Globe	American	Yes/Yes	0/0	65
1895-04-05	Atchison Daily Globe	American	Yes/Yes	0/5	83
1895-04-06	Atchison Daily Globe	American	Yes/Yes	0/0	156
1895-04-06	Atchison Daily Globe	American	Yes/Yes	0/1	201
1895-04-06	Atchison Daily Globe	American	Yes/Yes	0/0	94
1895-04-04	Atlanta Constitution	American	Yes/Yes	1/28	1170
1895-04-05	Atlanta Constitution	American	Yes/Yes	0/77	1004
1895-04-06	Atlanta Constitution	American	Yes/Yes	0/50	1180
1895-04-04	Boston Daily Advertiser	American	Yes/Yes	0/0	630
1895-04-05	Boston Daily Advertiser	American	Yes/Yes	0/0	404

informational unit in the news marketplace, the string algorithm could not simply operate on the paragraph level to be effective at detecting press networks (Slauter, 2012). The program, therefore, allows the algorithm to run both at the paragraph and document level; and for each individual news report, it indexes the number of document- and paragraph-level matches that have been identified (see Figure 1).

The Wilde Trials Web App also provides a variety of interfaces for viewing the reports, enabling different types of detailed textual comparison and analysis of text matching. First, there is a single-report view that collates at the paragraph level all matched text from other newspapers. This view also provides data about the report's document and paragraph-level matches, with links to the corresponding reports (see Figure 2).

Figure 2: Wilde Trials Web App single-report view



La Justice - Tuesday, May 07, 1895

Le procès Oscar Wilde

On mande de Londres, 5 mai.
D'apres les dernières informations, l'avocat d'Oscar Wilde aurait demandé au juge d'accepter, pour la mise en liberté de son client, une seule caution personnelle de 62,500 fr., au lieu d'y ajouter deux garanties de 31,250 francs chacune.

> LONDRES, 4 mai.--Par fil spécial. D'après les dernières informations, l'avocat d'Oscar Wilde aurait demandé au juge d'accepter, pour la mise en liberté de son client, une seule caution personnelle de 2,500 livres sterling, au lieu d'y ajouter deux garanties de 1,25O livres sterling chacune.
>
> Le Matin - Sunday, May 05, 1895 (80.3%)
> Compare two documents

Le juge fera connaitre sa decision lundi prochain.

Metadata

Publisher
La Justice
Date published
1895-05-07
Region
French
Word count
55

Document Matches

Indexed: Yes

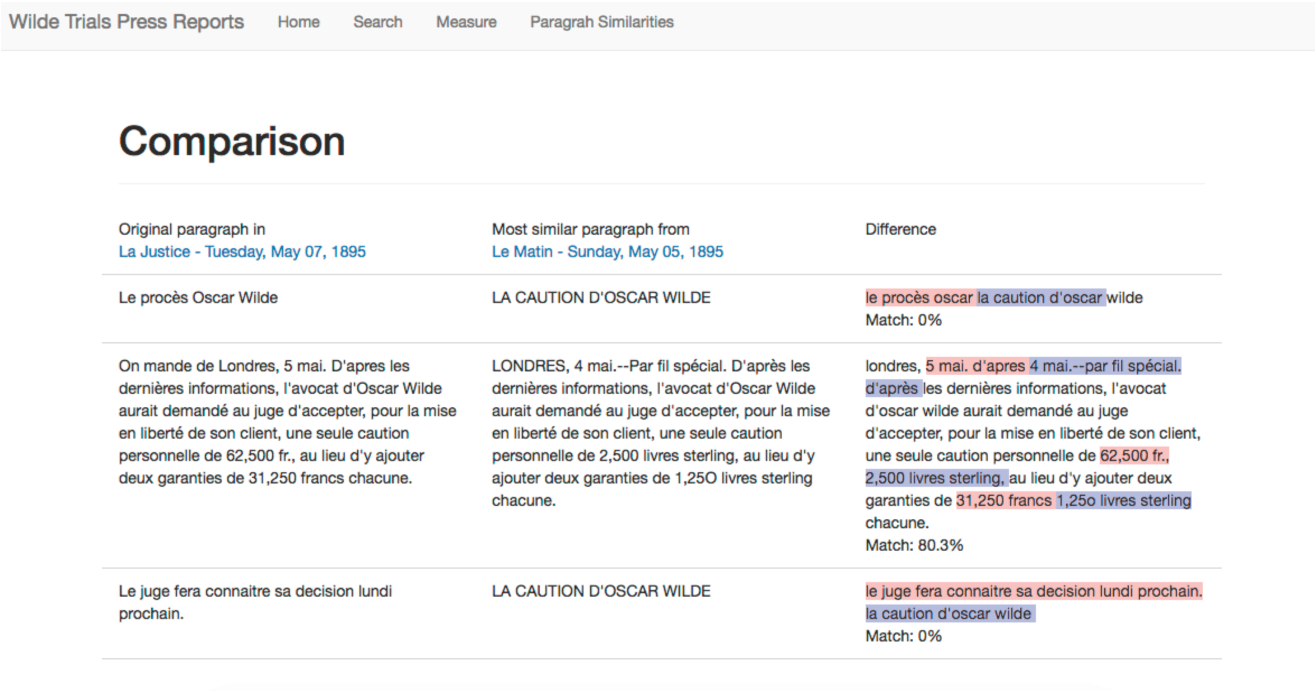
- Le Matin - Sunday, May 05, 1895 (64.0%)

Second, there is a side-by-side view of two reports. This interface shows three columns: the first is the base report, the second is the report to which it is being compared, and the third merges the two reports and features a double colour-code index highlighting deletions and additions made to the two documents (see Figure 3).

A third viewing interface lists all of the paragraph similarities across the entire database (currently a feature that is helpful for identifying the most highly diffused paragraphs during the Wilde trials) (see Figure 4).

Finally, there is a freestyle comparison interface that works as a text-comparison laboratory: it allows the user to select any two texts from within or outside the textbase, and it then runs different similarity metrics on the selected passages. The program's search function allows the user easily to grab these passages from anywhere in the textbase, and allows for high user interactivity with the textual material and quantified comparative analysis (see Figure 5).

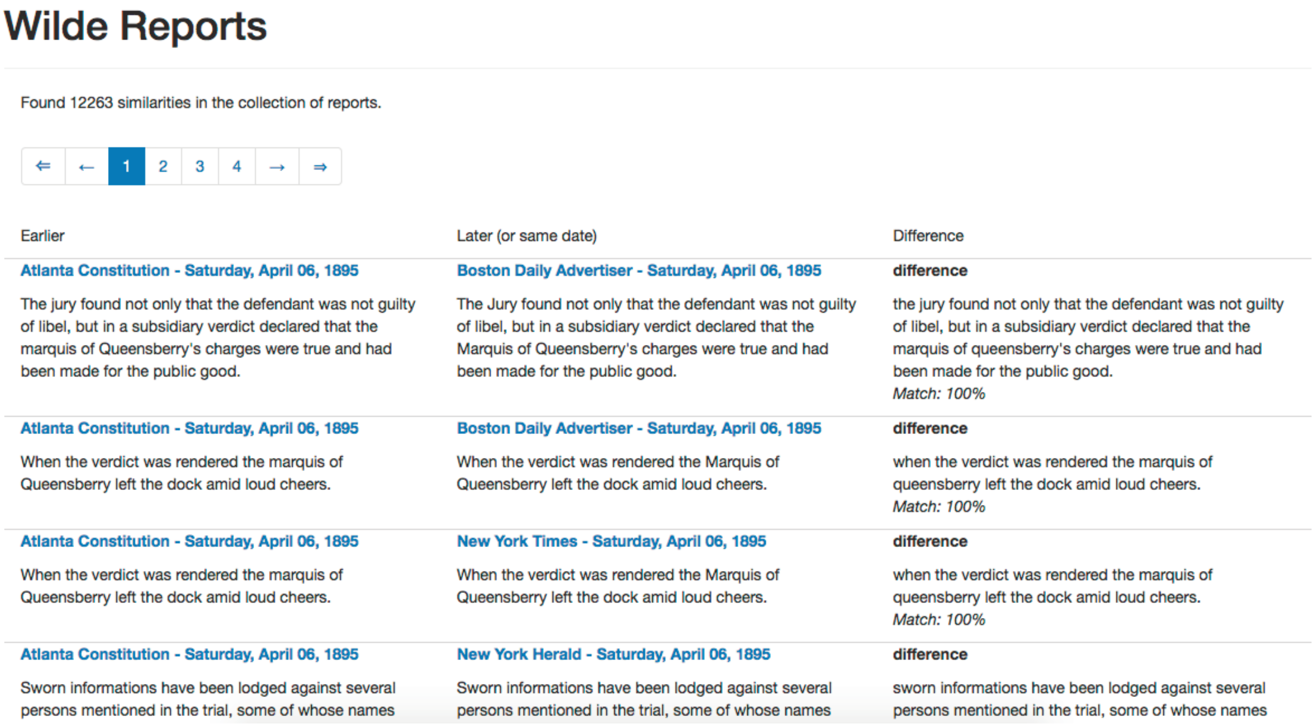
Figure 3: Wilde Trials Web App side-by side view of two reports



Comparison

Original paragraph in La Justice - Tuesday, May 07, 1895	Most similar paragraph from Le Matin - Sunday, May 05, 1895	Difference
Le procès Oscar Wilde	LA CAUTION D'OSCAR WILDE	le procès oscar la caution d'oscar wilde Match: 0%
On mande de Londres, 5 mai. D'apres les dernières informations, l'avocat d'Oscar Wilde aurait demandé au juge d'accepter, pour la mise en liberté de son client, une seule caution personnelle de 62,500 fr., au lieu d'y ajouter deux garanties de 31,250 francs chacune.	LONDRES, 4 mai.--Par fil spécial. D'après les dernières informations, l'avocat d'Oscar Wilde aurait demandé au juge d'accepter, pour la mise en liberté de son client, une seule caution personnelle de 2,500 livres sterling, au lieu d'y ajouter deux garanties de 1,25O livres sterling chacune.	londres, 5 mai. d'apres 4 mai.--par fil spécial. d'après les dernières informations, l'avocat d'oscar wilde aurait demandé au juge d'accepter, pour la mise en liberté de son client, une seule caution personnelle de 62,500 fr., 2,500 livres sterling, au lieu d'y ajouter deux garanties de 31,250 francs 1,25o livres sterling chacune. Match: 80.3%
Le juge fera connaitre sa decision lundi prochain.	LA CAUTION D'OSCAR WILDE	le juge fera connaitre sa decision lundi prochain. la caution d'oscar wilde Match: 0%

Figure 4: Wilde Trials Web App all paragraph similarities



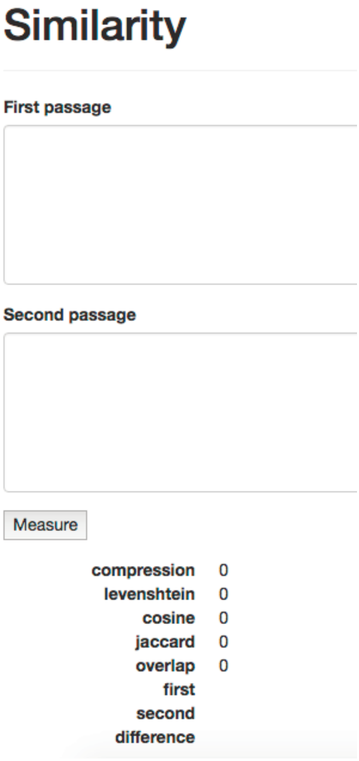
Wilde Reports

Found 12263 similarities in the collection of reports.

Earlier	Later (or same date)	Difference
Atlanta Constitution - Saturday, April 06, 1895 The jury found not only that the defendant was not guilty of libel, but in a subsidiary verdict declared that the marquis of Queensberry's charges were true and had been made for the public good.	Boston Daily Advertiser - Saturday, April 06, 1895 The Jury found not only that the defendant was not guilty of libel, but in a subsidiary verdict declared that the Marquis of Queensberry's charges were true and had been made for the public good.	**difference** the jury found not only that the defendant was not guilty of libel, but in a subsidiary verdict declared that the marquis of queensberry's charges were true and had been made for the public good. *Match: 100%*
Atlanta Constitution - Saturday, April 06, 1895 When the verdict was rendered the marquis of Queensberry left the dock amid loud cheers.	Boston Daily Advertiser - Saturday, April 06, 1895 When the verdict was rendered the Marquis of Queensberry left the dock amid loud cheers.	**difference** when the verdict was rendered the marquis of queensberry left the dock amid loud cheers. *Match: 100%*
Atlanta Constitution - Saturday, April 06, 1895 When the verdict was rendered the marquis of Queensberry left the dock amid loud cheers.	New York Times - Saturday, April 06, 1895 When the verdict was rendered the Marquis of Queensberry left the dock amid loud cheers.	**difference** when the verdict was rendered the marquis of queensberry left the dock amid loud cheers. *Match: 100%*
Atlanta Constitution - Saturday, April 06, 1895 Sworn informations have been lodged against several persons mentioned in the trial, some of whose names	New York Herald - Saturday, April 06, 1895 Sworn informations have been lodged against several persons mentioned in the trial, some of whose names	**difference** sworn informations have been lodged against several persons mentioned in the trial, some of whose names

With its full-text document database, text-sharing algorithm, and multiple interfaces for textual comparison and data display, the Web app is a powerful and versatile tool for both large-scale aggregate analysis and close detailed analysis of the international news coverage of the Wilde trials.

Figure 5: Wilde Trials Web App freestyle comparison interface



```
compression  0
levenshtein  0
     cosine  0
    jaccard  0
    overlap  0
      first
     second
 difference
```

Significant editorial labour and technical development have gone into creating the program's design and functionality, involving everyone on the research team. To start with the editorial work, the news reports are organized into directories. The top-level directory is the region where the report originated (America, France, Australia, Britain), and then there is one directory in each region for each paper from that region. Files follow strict naming conventions using the name and date of the publication. Automated scripts written in the Perl programming language check the file and directory names to report any inconsistencies and to ensure the names meet the strict standards.

Most of these reports start out as an OCR transcription (though some newspapers have been manually transcribed). Members of the research team open the text in a plain-text editor, such as TextWrangler, and convert the content to UTF-8 Unicode text. Ensuring that each step of the process uses the same character encoding is crucial to preventing problems down the line. The OCR files include all the text on the newspaper page. Someone on the research team will start by removing the unrelated text surrounding the report, an important part of the curatorial process that ensures that only news related to the trials is gathered and subject to quantification. Then an editor cleans the OCR by correcting spelling, word spacing, and punctuation. Finally, someone compares the corrected text against the report and inserts paragraph breaks as they appear in the original. Determining paragraph breaks requires some interpretation, as mentioned above, as the line and paragraph breaks are not always clear. An editorial policy is in place for the research team with instructions on paragraphing and diplomatic transcription. Once the files have been edited, they are processed into very simple, unstyled XHTML using a Perl script. The Perl script transforms paragraphs in the plain-text version to paragraphs in XHTML. The script also adds basic metadata for the region, paper title, publication date, and word count. This metadata is encoded following the Dublin Core Metadata Standards, generic terms used for describing a wide range of artefacts. The generated XHTML also contains markup for line and paragraph breaks.

The text files that have been transformed to XHTML are then loaded into eXist, an open source XML database, where they are indexed for similarity. XQuery modules process the documents and paragraphs for similarity in a few steps; they

1. List all the documents/paragraphs;
2. Normalize the documents/paragraphs by converting to lowercase and removing punctuation and extra space. This normalization reduces the influence of any transcription errors or stylistic differences in the papers; and
3. Compare each document/paragraph against every other document/paragraph using the Levenshtein distance string metric.

Levenshtein distance records the number of character edits (insertions or deletions) required to turn one string into another. Two documents that are similar will have very small Levenshtein distance. Turning the distance between two documents to a measure of similarity is straightforward:

similarity = 1 - distance / maximum length

The similarity between two documents is stored as a Dublin Core metadata element in the document's head. For paragraphs, the similarity is recorded as an HTML link at the end of the paragraph. These updates are done in eXist, using its XUpdate mechanism. Initial similarity indexing runs were quite slow, as comparing every document to every other document is a process (this big-O notation describes algorithms whose running time is directly proportional to the square of the input size). Some optimizations have therefore been implemented in the indexing code:

- Short strings do not need to be compared. In this system, strings shorter than 24 characters are ignored. This small optimization reduced the number of comparisons by ten percent.
- Similarity is a symmetric relationship. If documents A and B have a similarity of 75 percent, then documents B and A also have a similarity of 75 percent and there is no need to compare them. This cut the number of operations by half.
- The first paragraph in a document is the news report header. These headers are often similar – for example, "Wilde to Be Released on Bail" in *Daily Inter Ocean* on May 4, 1895, compared to "Wilde Released On Bail" in *Freeman's Journal* on May 8, 1895 – even if the documents containing them are not (Daily Inter Ocean, Freeman's Journal). The algorithm therefore does not run on headers.
- If the string lengths of documents A and B differ by more than 40 percent, then the documents must be less than 60 percent similar. We set our paragraph- and document-level match thresholds at 60 percent. This optimization provided the biggest improvement to the running time.

Without these optimizations, indexing the documents took eight hours. After optimizing the algorithm, indexing runtime takes about an hour.

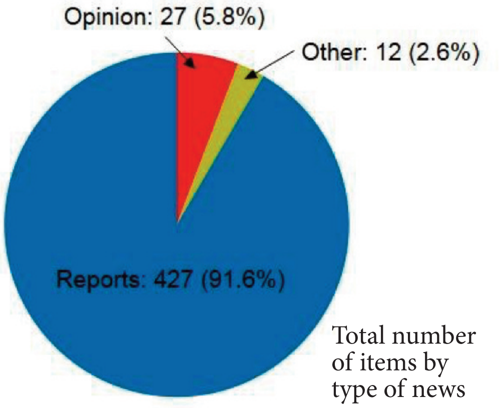
Other XQuery modules list the documents in the database and provide searching via the Lucene full-text engine, a search-function that is included with eXist. Documents are displayed in a Web browser via an XQuery module, which injects similar paragraphs from other documents in the output as blockquote XHTML elements.

After a few rounds of revision, over a roughly eight-month development period, the first stage of the program's design was complete and comparisons were run.

Early discoveries

The Wilde Trials Web App has already led to significant early discoveries about the Parisian press coverage of the trials. The program currently includes full-runs on the three trials from 18 different Parisian daily newspapers, most of which were gathered from the collection of periodicals digitized by Gallica. This corpus is comprehensive (though it does exclude a minority of newspapers that have not been digitized, as well as weekly and monthly periodicals), and it is also diverse, as it includes newspapers from across the political spectrum, as well as expatriate Parisian papers published in English. Leading scholarship on the French news reports on the trials has thus far focused on the opinion pieces (Erber, 1996). As important as these were for their commentary (on the British justice system, the moral judgement of art, the role of the press, and male sexual deviance), they accounted for only a small proportion of the total reporting on trials, less than six percent. Almost 92 percent of Parisian coverage was in the form of news reports (see Figure 6).

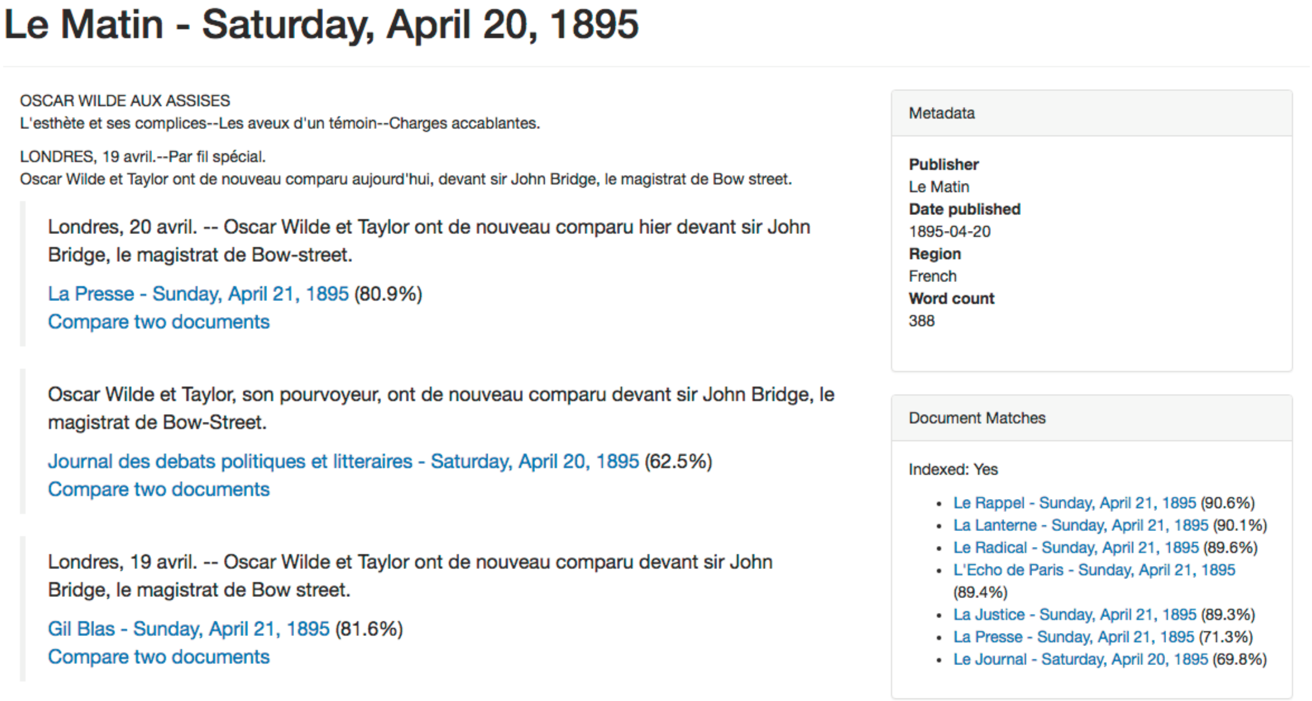
Figure 6: Types of news on the Wilde trials in the Parisian press



Total number of items by type of news

The Web app helps reveal the incredible extent to which the news circulating in Paris kiosks, offices, and cafés about Wilde's London trials sounded very much the same. Opening up almost any French-language news report on the trials within the program reveals numerous document-level matches with other French news reports. One report from the popular newspaper *Le Matin* shared as much as 90 percent identical text with reports from seven other Parisian newspapers (see Figure 7).

Figure 7: Document-level match data for Le Matin news report on April 20, 1895



Le Matin - Saturday, April 20, 1895

OSCAR WILDE AUX ASSISES
L'esthète et ses complices--Les aveux d'un témoin--Charges accablantes.

LONDRES, 19 avril.--Par fil spécial.
Oscar Wilde et Taylor ont de nouveau comparu aujourd'hui, devant sir John Bridge, le magistrat de Bow street.

> Londres, 20 avril. -- Oscar Wilde et Taylor ont de nouveau comparu hier devant sir John Bridge, le magistrat de Bow-street.
>
> La Presse - Sunday, April 21, 1895 (80.9%)
> Compare two documents

> Oscar Wilde et Taylor, son pourvoyeur, ont de nouveau comparu devant sir John Bridge, le magistrat de Bow-Street.
>
> Journal des debats politiques et litteraires - Saturday, April 20, 1895 (62.5%)
> Compare two documents

> Londres, 19 avril. -- Oscar Wilde et Taylor ont de nouveau comparu devant sir John Bridge, le magistrat de Bow street.
>
> Gil Blas - Sunday, April 21, 1895 (81.6%)
> Compare two documents

Metadata

Publisher
Le Matin
Date published
1895-04-20
Region
French
Word count
388

Document Matches

Indexed: Yes

- Le Rappel - Sunday, April 21, 1895 (90.6%)
- La Lanterne - Sunday, April 21, 1895 (90.1%)
- Le Radical - Sunday, April 21, 1895 (89.6%)
- L'Echo de Paris - Sunday, April 21, 1895 (89.4%)
- La Justice - Sunday, April 21, 1895 (89.3%)
- La Presse - Sunday, April 21, 1895 (71.3%)
- Le Journal - Saturday, April 20, 1895 (69.8%)

In fact, data derived from the Web app and visualized graphically using Tableau software reveals that 67 percent of *Le Matin*'s total reports published during the trials matched reports from other newspapers (see Figure 8).

Le Matin was by no means exceptional in this text-sharing practice: at least 20 percent of all reports published in the French-language papers shared content with other newspapers, and some newspapers were as high as 90 percent (see Figure 9).

Figure 8: Percentage of matching reports in *Le Matin* newspaper

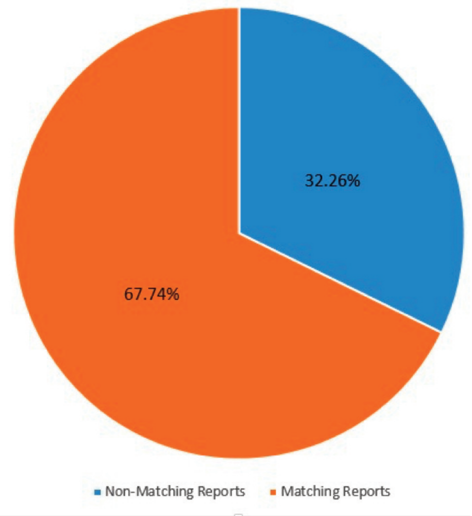
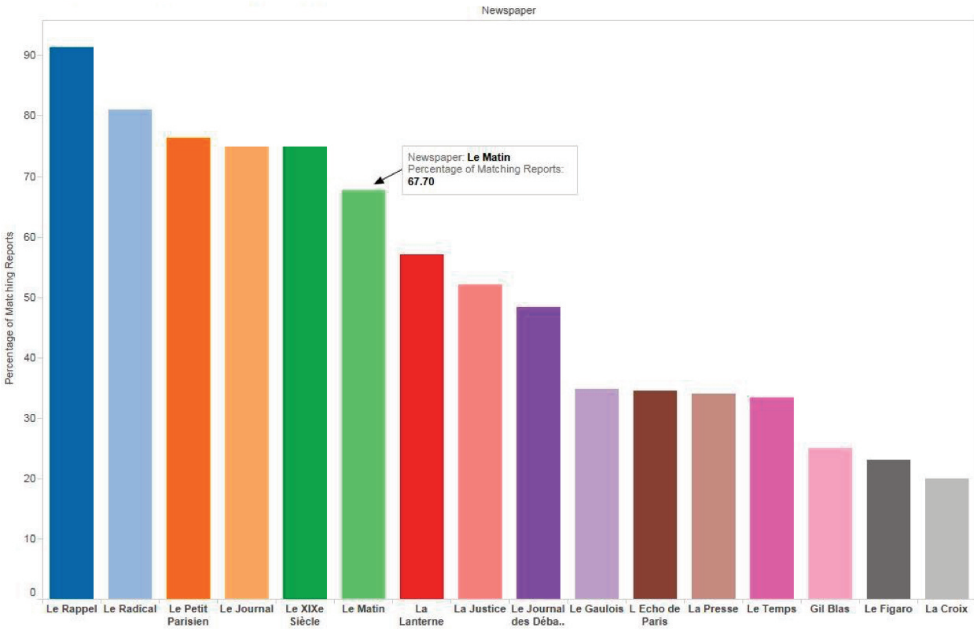
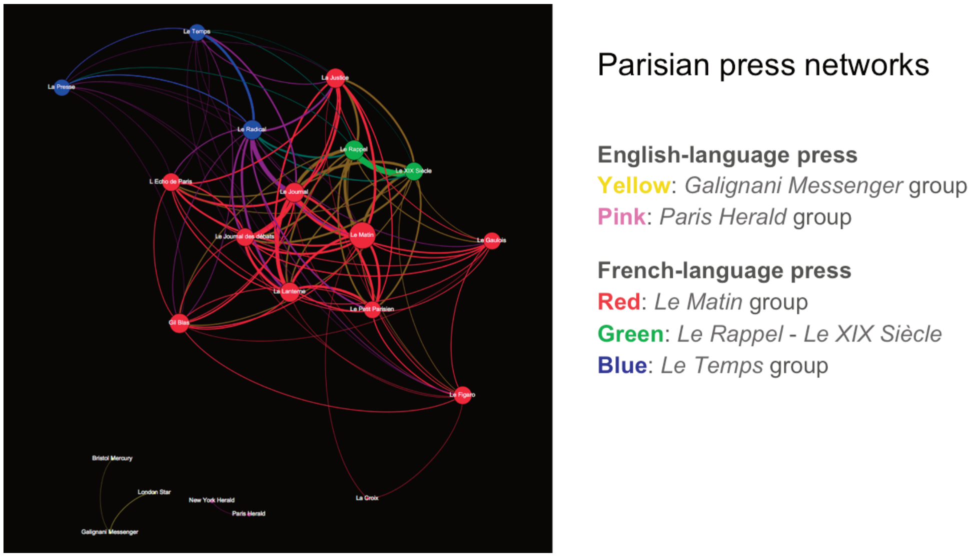


Figure 9: Percentage of matching reports on the Wilde trials in French newspapers



The document-level match data generated from the Web app also leads to new understanding of how these reports circulated within distinctive press networks. Document-level match data from the program uploaded into Gephi network visualization software generates a network graph that shows just how highly centralized the French-language press was (see Figure 10). Five groups of papers transmitted streams of news about the Wilde trials. The English-language expatriate press, which consisted of exactly two papers in 1895, made up two of these groups. They sourced their news by diverse means, so their news was different from the French-language dailies as well as from each other. The French-language daily press formed into another three groups of interconnected papers, or strongly connected components. The first mostly highly connected group consisted of two papers only (*Le XIXe Siècle* and *Le Rappel*) whose reports were very similar because they shared the same editorial offices. The two papers would eventually merge in 1899 (Albert, 1972). This group was also closely connected to another second group of seven papers (*Le Matin*, *Le Journal*, *Le Petit Parisien*, *L'Echo de Paris*, *La Lanterne*, *La Justice*, and *Le Gaulois*). Both of these groups reprinted many of the same international telegrams from the international press agency Havas, one of the big four global news agencies of the period (Barth, 2014; Silberstein-Loeb, 2014). The third group of French papers (a group of five, including *Le Temps*, *Le Figaro*, *Le Gaulois, La Presse*, and *La Croix*) was more loosely connected to these other two groups, as well as to each other. Newspapers like *Le Temps* and *Le Figaro* also published Havas telegrams about the trials, which explains their textual connectivity to the other two groups, but they supplemented these telegrams with other news-gathering methods and invested in good editing and writing that resulted in more diversified reporting on Wilde's trials (Albert, 1972). The Web app thus generates match data that provides the first quantitative picture of newsmaking in Paris. It is also generating data-driven reception analysis of the trial coverage in terms of volume, velocity, veracity, and networked activity. Such analysis might also help us understand whether the "complex contagion" principle from sociology might apply to the coverage of the Wilde trials (Centola & Macy, 2007): in other words, was repeated exposure to the same news content integral to the adoption of such a highly controversial topic?

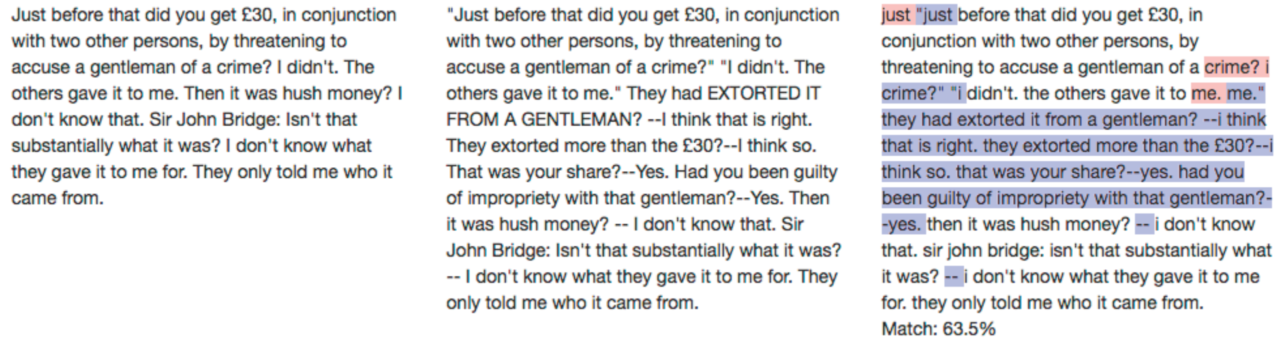
Figure 10: Draft text-sharing graph of Wilde trials news reports in French daily newspapers; colour scheme identifies the five different news networks



Along with a quantitative picture of the Parisian news coverage, the Web app also helps detect divergent news and different local news standards. The interface for side-by-side textual comparison, for example, has furthered understanding of how the expatriate newspaper *Galignani Messenger* gathered and diffused news about the trials from across the British Channel. Where one would have expected the Paris-based paper to be more explicit than the British press, it repeatedly censored the sexual information published in the London newspapers from which it sourced its news. In its April 12 report (*Galignani Messenger*), it omitted compromising testimony regarding homosexual blackmail, which both its London source papers included. The British newspapers *Star* and *Reynolds's Newspaper* transcribed an exchange during which one of the witnesses in the trial, an ex-valet named Charles Parker, admitted to "impropriety" with an unnamed gentleman who was the target of a blackmailing scheme (*The London Star, Reynolds's Newspaper*). Although the *Galignani Messenger* report shared essentially the same news report, it excluded this remarkable admission, as well as the fuller questioning about extortion (see Figure 11). This omission was not simply an instance of trimming, but a strategic cut that minimized disclosure about sexual blackmail and the larger cultural process of the trials that linked sodomy to blackmail (Bristow, 2016). That the *Galignani Messenger* moderately censored some of the sexual information from the London papers invites questions about the extra-national reach of Britain's journalistic convention of silence around homosexual matters (Cocks, 2003; Powell, 2009), and overturns expectations of finding sexual tolerance and expressive freedom in a newspaper based in the City of Light.

These early findings about the Parisian press indicate the kinds of qualitative and quantitative analysis of the international news coverage of the trials facilitated by the Web app.

Figure 11: Evidence of censorship in the *Galignani Messenger*. *Galignani Messenger* report from April 12, 1895 (left), compared to a *London Star* report from April 11, 1895 (right).



Just before that did you get £30, in conjunction with two other persons, by threatening to accuse a gentleman of a crime? I didn't. The others gave it to me. Then it was hush money? I don't know that. Sir John Bridge: Isn't that substantially what it was? I don't know what they gave it to me for. They only told me who it came from.

"Just before that did you get £30, in conjunction with two other persons, by threatening to accuse a gentleman of a crime?" "I didn't. The others gave it to me." They had EXTORTED IT FROM A GENTLEMAN? --I think that is right. They extorted more than the £30?--I think so. That was your share?--Yes. Had you been guilty of impropriety with that gentleman?--Yes. Then it was hush money? -- I don't know that. Sir John Bridge: Isn't that substantially what it was? -- I don't know what they gave it to me for. They only told me who it came from.

just "just before that did you get £30, in conjunction with two other persons, by threatening to accuse a gentleman of a crime? i crime?" "i didn't. the others gave it to me. me." they had extorted it from a gentleman? --i think that is right. they extorted more than the £30?--i think so. that was your share?--yes. had you been guilty of impropriety with that gentleman?--yes. then it was hush money? -- i don't know that. sir john bridge: isn't that substantially what it was? -- i don't know what they gave it to me for. they only told me who it came from.
Match: 63.5%

Going forward

The Wilde Trials Web App is the only online curation of international news reports on the Wilde trials. Unlike Ryan Cordell's Viral Texts Project, which examines how, why, and what kinds of texts were reused in nineteenth-century American newspapers, the Wilde Trials Web App features a corpus of texts that is clearly defined around a singular, major news event. These texts are corrected, and meant to aid both reading and analysis. When the Web app is publicly launched, projected for spring 2017, it will include a reading interface for reading the collection of the reports, as well as the different analytical interfaces. As we prepare for the launch, we will also be developing some of its functions.

One aspect of the program in development is its text-sharing detection algorithm. The Web app's tools are currently based on Levenshtein distance, which measures the number of edits required to turn one passage of text into another. Other measures of similarity include cosine (which shows similarity based on word frequency without regard to word order) and the Jaccard index (which shows similarity based on words common to two passages). We expect that the different measures of these string metrics will reveal different patterns of copying and text sharing in the corpus, which could bear on our analysis. It may be possible to refine the cosine word-frequency similarity by using N-grams (sequences of two, three, or more words) to detect plagiarism. The use of natural language processing (NLP) tools could also be used to reduce inflected or derived words to their root word stem. The program's algorithms consider "examination" and "examined" to be different words even though they are both derived from "exam." It may be possible to detect more instances of copying by applying stemming or lemmatization to the text before processing. We are also exploring machine-generated translation for detecting text sharing from English into French.

In addition to improving our text-sharing algorithm, we are adding visual analytics to facilitate our understanding of news coverage in relation to the international news structures that shaped it. We plan to incorporate mapping applications to visualize our data on the volume, velocity, and veracity of the reports in different regions of the world. We have also been working on incorporating a network analysis application into the Web app, which will organize text-sharing data into networked relationships, displaying the different flows of news on the Wilde trials. Incorporating a network graph application into the program will obviate the need to migrate the data over to other software, as has been our practice thus far. These refinements to the Web app should lead to richer and more robust findings on news sharing and news flows, and increase our understanding of the geopolitical and market forces behind the journalistic activity. The program will also offer the ability to extract the metadata into comma-separated values (CSV) files for processing outside the research environment.

We designed the Web app to help answer a computational question about shared news content in the reporting of the Wilde trials, but as we have developed and used it, its potential utility beyond its original design has emerged. The program might provide the architecture for other large text-matching databases equipped with "similarity tools to discover shared passages, borrowings, plagiarisms, and other forms of text recycling" (Cooney, Roe, & Olsen, 2014, para. 65). The algorithms and software are sufficiently generic that any simple XHTML data set could be used. As it was developed in the context of plagiarism software, it could be offered as an open access alternative to currently available commercial products, such as Turnitin. It could also be used to compare other sets of documents, from the past to the present day, to investigate text sharing or viral news events relevant for textual criticism, attribution studies, or stylistic analysis.

The Web app has also generated an unexpected spinoff. Our experimentation with the free-measure interface, which highlights revision text, revealed its potential use for writing and revision, and has led to further collaboration designing a Web-based digital writing revision tool. Text sharing and remixing are, and have been, central to our cultural literacy. The text-comparison tools that we are designing are charting ways forward for interactive qualitative and quantitative study of this fascinating textual phenomenon, and for tools that will assist with its perpetuation.

Websites

CollateX	http://collatex.net/
Dublin Core Metadata Standards	http://dublincore.org/documents/dc-html/
Europeana	http://www.europeana.eu/portal/en
Gale-Cengage	http://www.gale.com
Gallica	http://gallica.bnf.fr/
Gephi	https://gephi.org/
eXistdb	http://exist-db.org/exist/apps/homepage/index.html
Juxta Commons	http://juxtacommons.org/
Mergely	http://www.mergely.com/
Tableau	http://www.tableau.com/
Trove	http://trove.nla.gov.au/newspaper/
Turnitin	http://turnitin.com/
Viral Texts	https://viraltexts.org/

References

Albert, Pierre. (1972). La presse française de 1871 à 1940. In Claude Bellanger, Jacques Godechot, Pierre Guiral, & Fernand Terrou (Eds.), *Histoire générale de la presse française. Tome III: De 1871 à 1940* (pp. 135-622). Paris, FR: Presses Universitaires de France.

Barth, Volker. (2014). The formation of global news agencies, 1859-1914. In W. Boyd Rayward (Ed.), *Information beyond borders: International cultural and intellectual exchange in the Belle Époque* (pp. 35-47). London, UK: Ashgate.

Bristow, Joseph. (2016). The blackmailer and the sodomite: Oscar Wilde on trial. *Feminist Theory, 17*(1), 41-62. URL: http://fty.sagepub.com.proxy.lib.sfu.ca/content/17/1/41 [June 1, 2016].

Centola, Damon, & Macy Michael. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology, 113*(3), 702-734. URL: http://www.jstor.org.proxy.lib.sfu.ca/stable/10.1086/521848 [June 1, 2016].

Cocks, Harry G. (2003). *Nameless offences: Homosexual desire in the 19th century*. London, UK: I.B. Tauris.

Cohen, Ed. (1993). *Talk on the Wilde side: Toward a genealogy of a discourse on male sexualities*. New York, NY: Routledge.

Cooney, Charles, Roe, Glenn, & Olsen, Mark. (2014). The notion of the textbase: Design and use of textbases in the humanities. *Literary Studies in the Digital Age.* URL: https://dlsanthology.commons.mla.org/the-notion-of-the-textbase/ [June 1, 2016].

Dekker, Ronald H., van Hulle, Dirk, Middell, Gregor, Neyt, Vincent, & van Zundert, Joris. (2014). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities, 30*(3), 452-470. URL: http://dx.doi.org.proxy.lib.sfu.ca/10.1093/llc/fqu007 [June 1, 2016].

Erber, Nancy. (1996). The French trials of Oscar Wilde. *Journal of the History of Sexuality, 6*(4), 549-588. URL: http://www.jstor.org.proxy.lib.sfu.ca/stable/4617221 [June 1, 2016]

Foldy, Michael. (1997). *The trials of Oscar Wilde: Deviance, morality, and late Victorian society.* New Haven, CT: Yale University Press.

Fotheringham, Richard. (2003). Exiled to the colonies: 'Oscar Wilde' in Australia, 1895–1897. *Nineteenth Century Theatre and Film, 30*(2), 55-68. URL: http://dx.doi.org.proxy.lib.sfu.ca/10.7227/NCTF.30.2.4 [June 1, 2016].

Ivory, Yvonne. (2012). "Aus anlass eines sensationsprozesses": The Oscar Wilde scandal in the German press. *Seminar: A Journal of Germanic Studies, 48*(2), 218-239. URL: http://muse.jhu.edu.proxy.lib.sfu.ca/article/475093 [June 1, 2016].

McGann, Jerome. (1991). *The textual condition*. Princeton, NJ: Princeton University Press.

No author. (1895, April 12). The society scandal. Police court proceedings. Oscar Wilde again in the dock. Weaving the web closer. The trip to Paris. Application for bail. *Galignani Messenger*, n. p.

No author. (1895, May 4). Wilde to be released on bail. *The Daily Inter Ocean*, p. 3.

No author. (1895, May 8). Oscar Wilde released on bail. *Freeman's Journal and Daily Commercial Advertiser*, p. 6.

No author. (1895, April 14). Oscar and Taylor. Again in the dock. "An Easter egg." Three writs and a cheque book. What was found at Taylor's house. Further disclosures. *Reynolds's Newspaper*, n. p.

No author. (1895, April 11). Oscar Wilde brought up at Bow-St this morning. Sir Edward Clarke again takes up Wilde's case, but declines to cross-examine the witnesses already called. *The London Star*, n.p.

Powell, Kerry. (2009). *Acting Wilde: Victorian sexuality, theatre, Oscar Wilde*. Cambridge: Cambridge University Press.

Robinson, Greg. (2015). Whispers of the unspeakable: New York and Montreal newspaper coverage of the Oscar Wilde trials in 1895. *Journal of Transnational American Studies, 6*(1), 1-18. URL: http://escholarship.org/uc/item/74n7c590 [June 1, 2016].

Ross, Stephen, & Sayers, Jentery. (2014). Modernism meets digital humanities. *Literature Compass, 11*(9), 625-633. URL: http://onlinelibrary.wiley.com.proxy.lib.sfu.ca/doi/10.1111/lic3.12174/full [June 1, 2016].

Sarony, Napoleon. (1882). URL: http://www.loc.gov/pictures/item/98519699/

Shahabi, Mitra. (2012). Comparing three plagiarism tools (Ferret, Sherlock, and Turnitin). *International Journal of Computational Linguistics, 3*(1), 53-66. URL: http://www.cscjournals.org/library/manuscriptinfo.php?mc=IJCL-33#MCAI [June 1, 2016].

Silberstein-Loeb, Jonathan. (2014). *The international distribution of the news: The Associated Press, Press Association, and Reuters, 1848-1947*. Cambridge, UK: Cambridge University Press.

Slauter, Will. (2012). The paragaph as information technology. *Annales. Histoire, Sciences Sociales, 2,* 253-278. URL: http://www.cairn.info/revue-annales-2012-2-page-253.htm [June 1, 2016].

Walshe, Eibhear. (2005). The first gay Irishman? Ireland and the Wilde trials. *Aire/Ireland, 40*(3), 38-57. URL: http://muse.jhu.edu.proxy.lib.sfu.ca/article/189970 [June 1, 2016].

Wan, Marco. (2006). From the rack to the press: Representation of the Oscar Wilde trials in the French newspaper *Le Temps. Law and Literature, 18*(1), 47-67. URL: http://www.jstor.org.proxy.lib.sfu.ca/stable/10.1525/lal.2006.18.1.47 [June 1, 2016].