UVicSPACE: Research & Learning Repository

University of Victoria

_____

Implementing New Knowledge Environments (INKE)

Publications

_____

Meditating on a Mandala in Class: Studying Shakespeare's Plays with a Visual Exploration Tool for XML Texts

Stéfan Sinclair, Stan Ruecker, Sandra Gabriele, Matt Patey, Matt Gooding, Chris Vitas, & Bartosz Bajer

2011

With permission from _Media : Culture : Pedagogy_

This article was originally published at:

01: Sinclair, et. al

Meditating on a Mandala in Class: Studying Shakespeare's Plays with a Visual Exploration Tool for XML Texts

Stéfan Sinclair, Stan Ruecker, Sandra Gabriele, Matt Patey, Matt Gooding, Chris Vitas, & Bartosz Bajer

Abstract

In this paper we describe the Mandala Browser (mandala.humviz.org) as a born-digital resource for use in the classroom. We provide example classroom exercises for studying the plays of Shakespeare, which provide on the one hand a simple means of examining speeches within a single play (our example uses *Romeo and Juliet*), and on the other hand a comparison between plays (e.g. all the tragedies). Finally, we provide some further context and resources for enabling what we call digital reading: a subset of text analysis oriented toward searching, browsing and reading text, without requiring more advanced knowledge of statistics and computational methods.

Introduction

It is estimated that humans will produce an unfathomable 1 zettabytes (21 zeroes) of digital information in 2010 (Gantz et al., 2008). However, the overabundance of data is not a new phenomenon of the information age. Vannevar Bush, the science advisor to the President of the USA during World War II, was already, in 1945, lamenting the disjoint between the quantity of information being produced and our tools for managing and finding that information: "The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present day interests, but rather that publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships" (Bush 1945). In the same article Bush describes the conceptual Memex machine that would inspire hypertext theorists and researchers and the development of the web. Although we have less ambitious goals, our motivations are similar to those of Bush: we are interested in creating tools that facilitate finding and understanding information in a mass of data. In particular, the Mandala Browser (mandala.humviz.org) described in this article is an attempt to balance power and user-friendliness in the design of a generalized tool for exploring XML files (Figure 1).
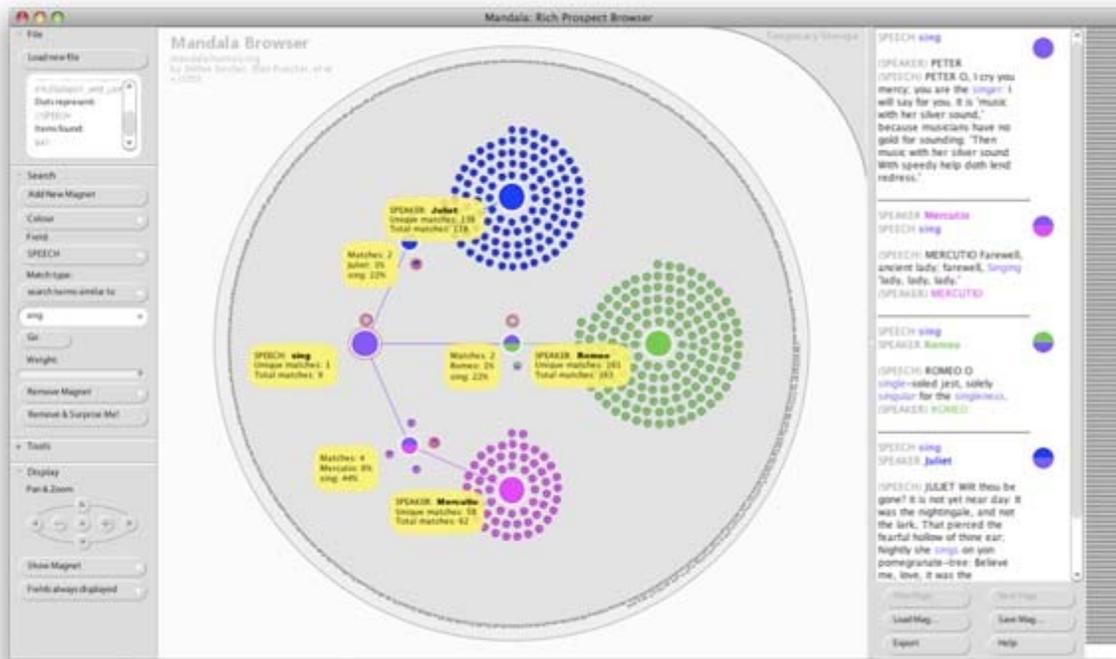
Figure 1: This Mandala shows the speeches in *Romeo and Juliet* as dots, some of which (the grey ones) are around the periphery, while others (the ones in colour) have been attracted to magnets in the interior of the Mandala. Here the student has created magnets to attract all speeches by the characters Romeo, Juliet, and Mercutio, with another magnet for any speech containing "sing." Note that the search choice "words similar to" means that the magnet has attracted speeches with "sing," "singer," and "singing," as well as "single" and "singular."

Based on a concept originally proposed by Oksana Cheypesh (see Cheypesh et al., 2006), the Mandala is a circular interface that allows people to dynamically construct visual Boolean queries of any XML-encoded text file or text collection. Each text is divided into subsections that appear around the periphery of the Mandala as dots. The divisions can be made at any points where an XML tag provides a possible subdivision. For plays, a natural unit is the speech, while in prose, it is often useful to work with dots as paragraphs (though one could work on words or chapters or any other unit defined by the XML markup). The user of the system can read the text behind each dot by clicking on it. The flexibility of the Mandala as a browser of structured text is a deliberate attempt to respond to two major challenges that have been articulated by a variety of scholars in the digital humanities: how do we make good use of the growing numbers of digital texts, and how does this use differ from the conventional methods of closely studying single texts? The consensus seems to be growing that we have reached a point of critical mass, where new methods of analyzing and discussing texts will become commonplace. For example, Moretti (2005) has adopted an approach that he calls "distant reading," where he observes patterns of change over time of phenomena such as the length of book titles–that is, he is using digital texts and computing power to answer hypotheses about the history of the book. Ramsay (2003), Unsworth (2005), Crane (2006), and Manovich (2009), on the other hand, suggest that tools for manipulating and visualizing text will provide students and other researchers with new ways, not of answering existing hypotheses, but instead of conceiving and pursuing new hypotheses. On a slightly different trajectory, Anderson (2008) and Halevy et al. (2009) propose that the age of hypotheses is over, and that instead of conceiving mental models for testing, researchers with access to enough data can instead directly observe and report on the patterns that emerge. The Mandala Browser does not commit to any of these perspectives in particular, but generally seeks to enable and enhance a range of reading and analytic practices. For classroom use, the Mandala can serve as the basis for formulating hypotheses either about individual texts or collections of texts. For example, in studying Shakespeare, the teacher might have the students work in depth with a single play, to support literary practices of close reading, or across multiple plays to encourage thinking about patterns

across subsets such as the history plays, comedies, or tragedies.

Studying *Romeo and Juliet* with the Mandala

The Mandala can be used to address both comparative questions and content questions. How much can be done depends in part on the XML encoding. Our examples expect that the play will have the following information marked in the XML: Acts, Scenes, Speakers (or Characters), and Speeches (because the plays are initially in XML, it is relatively straightforward to transform them to a different structure that may work better with the Mandala Browser; the original XML documents are from the WordHoard collection at Northwestern University.

The following are examples of questions that the Mandala Browser can help study:

<u>Comparative questions:</u>

- Which act contains the most speeches?
- Which act contains the most speeches by Romeo, by Juliet?
- Which character has the most speeches?
- Who says the following words the most, and what does this imply?

  o Love

  o Hate

  o War

- Who refers most to the following concepts:

  o Family, families, family relationships (mother, father, sister, brother)

  <u>Content questions:</u>

- In what act and scene does Mercutio stop speaking? Why?

For the purposes of providing a step-by-step example, we will look at use of the word "love." In working with the Mandala, a student begins by opening a play and indicating that the dots should represent speeches (Figures 2 and 3). Note that in Figure 3, the panel in the top left of the screen can be scrolled down to reveal that there are a total of 841 speeches in the play, which means that the students using the Mandala will have 841 dots to work with.
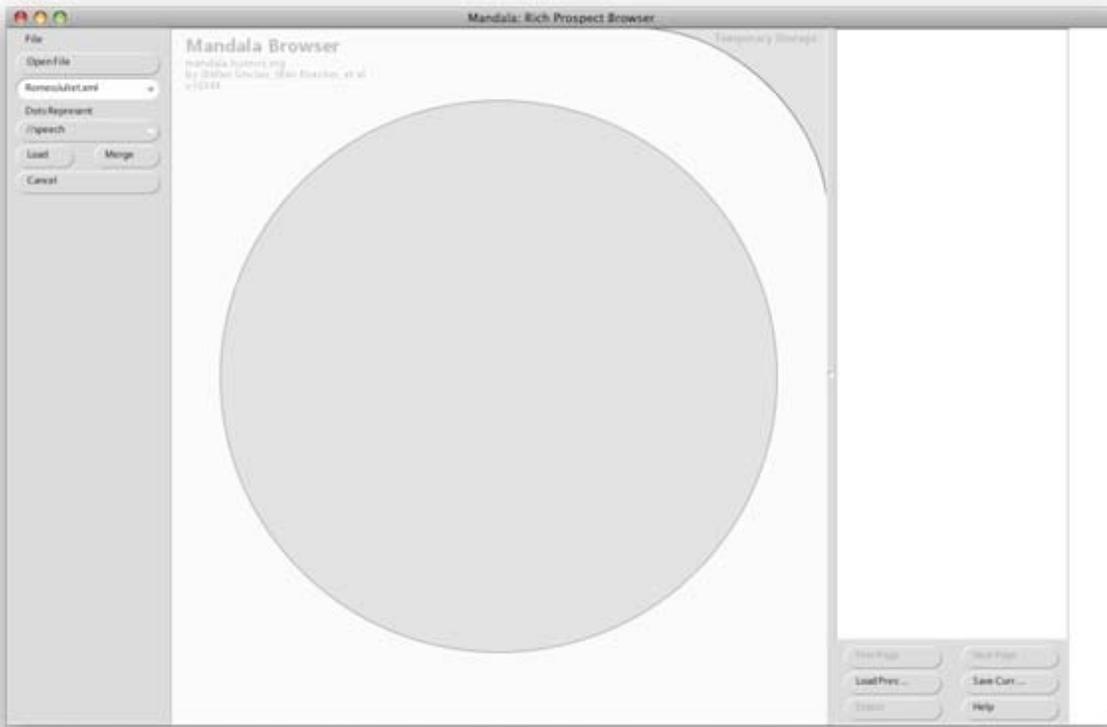
Figure 2: The Mandala has just been launched but has no document displayed.
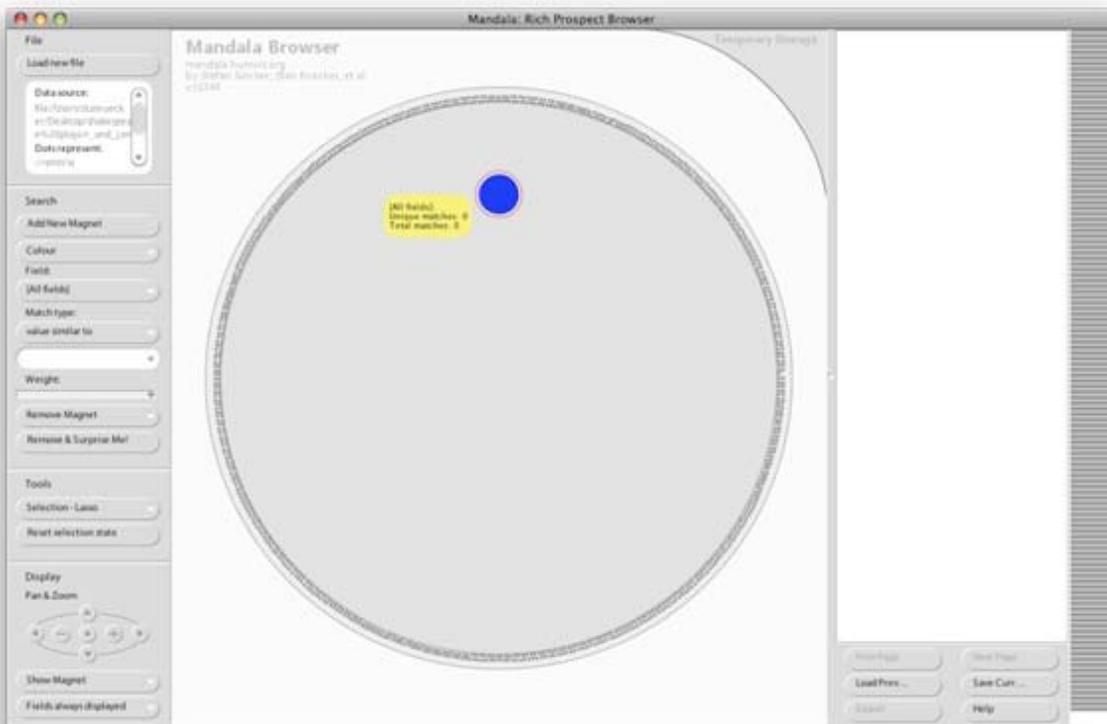

Figure 3: Here, the student has opened an XML-encoded version of *Romeo and Juliet* and indicated that each dot will represent one speech. Since there is only one blank magnet, all the dots appear as small, grey circles placed around the periphery.

The next step involves producing a series of nodes or "magnets" that attract the dots from the periphery into the interior of the Mandala. Figure 4 shows two magnets: one for all the speeches by Romeo and the other for all the speeches by Juliet.
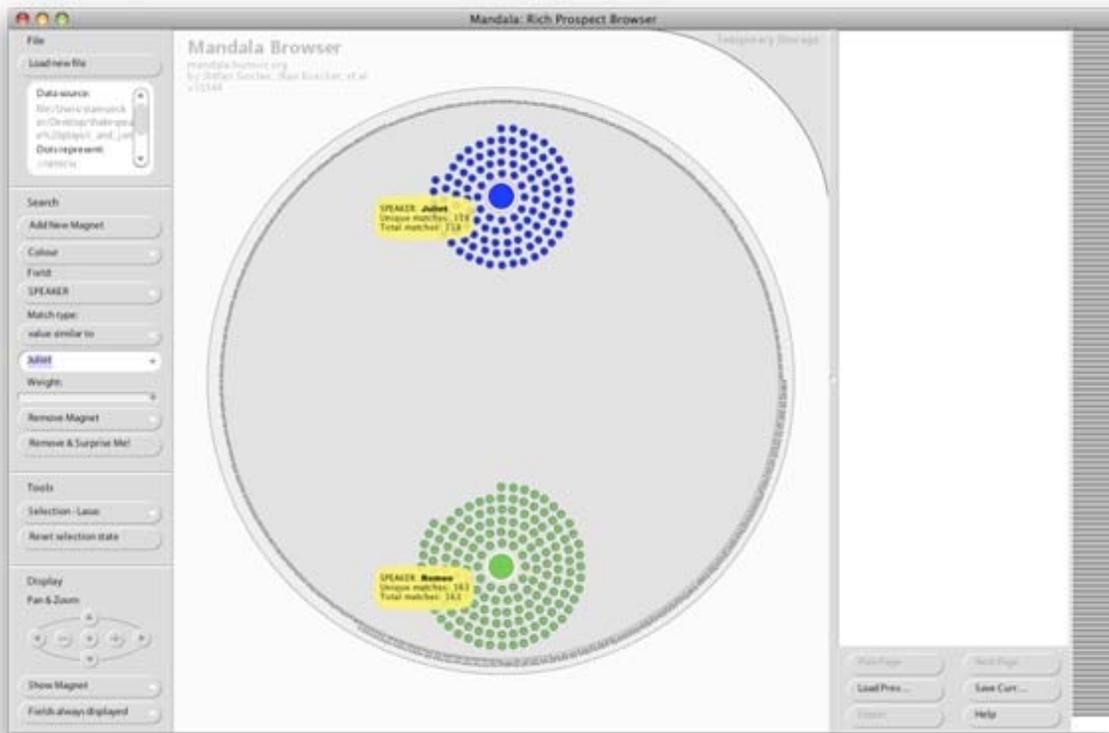
Figure 4: The student has created two magnets. Juliet's speeches appear at the top and Romeo's speeches at the bottom. Note that the magnets could be created in any order. Their colours can also be modified by the user.

At this stage it is possible to see that Romeo has quite a few more speeches than Juliet has, with 163 for Romeo and 118 for Juliet (the speech counts are indicated in the yellow label near each magnet). The next step is to begin to examine which of the two characters says what key words, how often those words are spoken by which character, and how often the words occur with respect to the total number of speeches for each character. Figure 5 shows the results for the word "love," where Romeo says "love" more often than Juliet says it. Of the total number of speeches where the word "love" is used (107 total), 34% of them are spoken by Romeo (37 by him out of 107 total) and 21% (23 out of 107) by Juliet. That is, one-third of all speeches in the play containing the word love are spoken by Romeo, and only one-fifth of the speeches using "love" are spoken by Juliet. However, these numbers don't take into account the fact that Romeo has quite a few more speeches than Juliet does. In terms of frequency, it is also true that the ratio of Romeo's total speeches including the word "love" is greater than the ratio of Juliet's speeches, but not by much. Romeo has 163 speeches in total, and he says "love" in 37 of them, or 22% of the time. Juliet has a total of 118 speeches, and she says "love" in 23 of them, or 19% of the time. Students might then hypothesize as to what this variance might suggest.
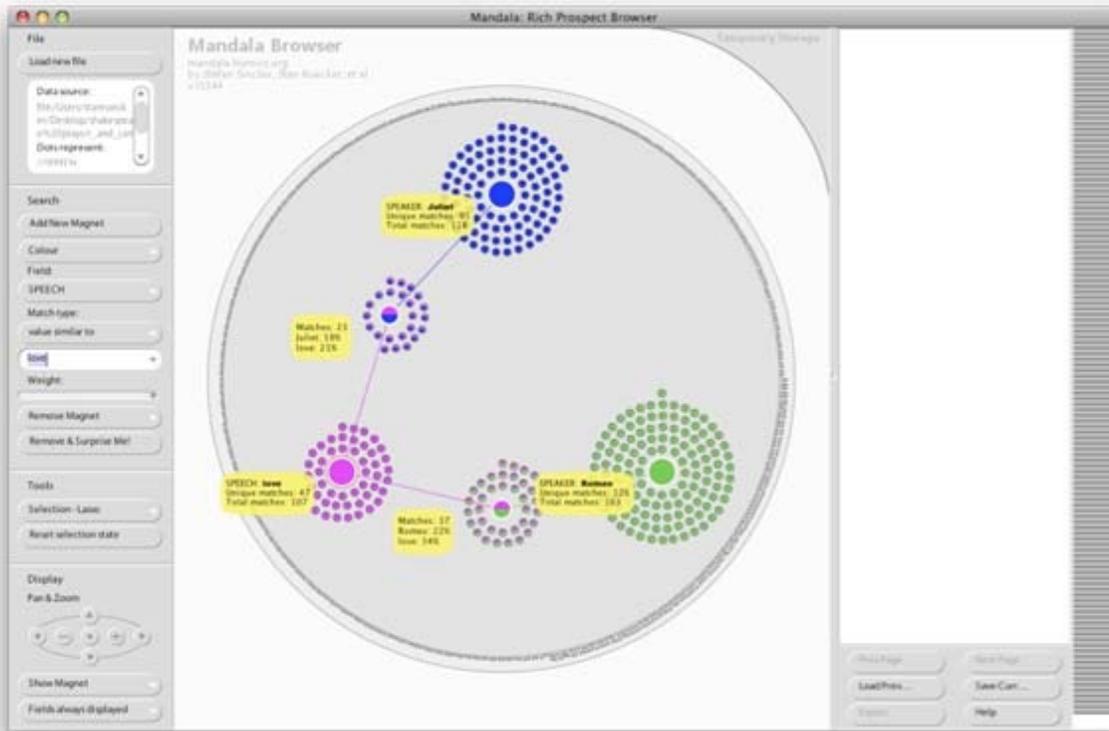
Figure 5: Romeo says "love" in more speeches than Juliet does, but then he also has more speeches. As a ratio of each of their total speeches, they say "love" almost the same percentage of the time.

Between them, *Romeo and Juliet* are responsible for 60 out of 107 speeches where someone says "love," or 56% of all speeches where someone says "love." In order to see who is responsible for the remaining 47 speeches, it would be possible to look individually through each of the texts represented by the dots, by using the lasso tool to select all those 47 dots for display in the reading panel on the right (Figure 6).
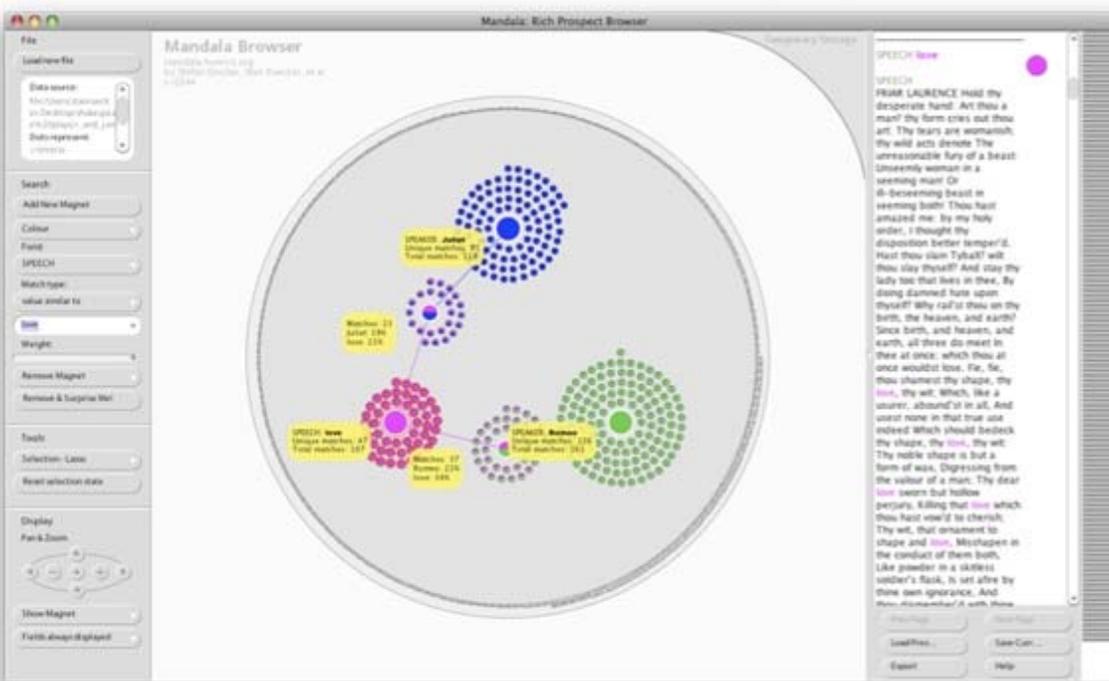

Figure 6: The student here is examining the texts of the 47 speeches where someone other than Romeo or Juliet says the word "love." The text associated with selected dots appears in the right-hand reading panel.

Alternatively, it is possible to see one magnet for each of the speakers in the play. The Mandala does this automatically when the user chooses the field "Speaker" and the value "[All Values]." In the case of *Romeo and Juliet*, the result is a rather large set of 35 magnets—one for each speaker in the play (Figure 7).

Figure 7: This Mandala shows one magnet for each speaker in the play. In this screenshot, the magnet labels have been temporarily turned off in order to make the magnets easier to see.

Once the Mandala has provided one magnet per speaker, we can add a magnet to see who says "love." It appears in Figure 8 as a dark greenish dot near the apex, with many subsets connected to it. After *Romeo and Juliet*, the characters who use the word "love" the most are Lady Montagu, Benvolio, and Paris. Somewhat surprisingly, given that she has a great many speeches in the play (in fact, 90), the Nurse only says "love" four times, and Friar John, who is similarly instrumental in the tragedy, only says it in five speeches out of his total of 59 speeches.
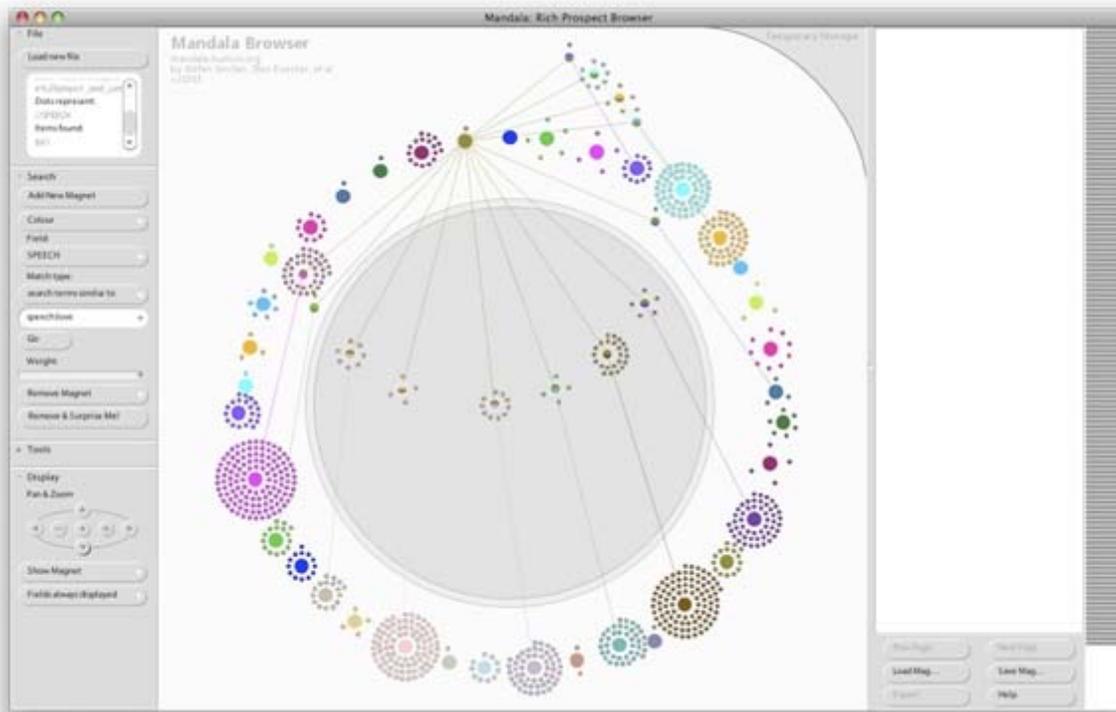
Figure 8: For the 35 characters in the play *Romeo and Juliet*, who says "love"?

As a next step, it is now possible to manually divide the speakers into the genders male and female in order to address the following question: is Romeo's use of the word "love" characteristic of the other men in the play, and is Juliet's use reflected by the other women? In a similar vein, we can think of the characters being divided by age or social status. The answer appears to be that use of the word "love" is less connected with any of these factors than it is with patterns of association (if the XML had contained gender, age, or social status markers for each character, this would have been even easier to investigate). Romeo's mother and his friends Benvolio and Mercutio are both on the high-frequency list, and so is his rival Paris. Juliet, on the other hand, stands out as an anomaly among the people she knows. One implication worth pursuing further is that Romeo and Paris may actually share the use of conventional rhetorical language – suggesting among other things that Juliet's parents may not have been so far off the mark in proposing Paris as her husband. Further investigation is also possible concerning other key words in the play, such as "death." It is also often useful to study how and when other characters use the name of a principal character. For instance, the name "Romeo" is used in 82 speeches by other characters, and "Juliet" is used in 40 speeches (13 of them by Romeo). In the same way that we have examined the word "love" in *Romeo and Juliet*, it is equally possible to investigate other key terms in other plays. One might, for instance, spend time on the word "blood" in *Macbeth*, or "citizen" in *Julius Caesar*. In order to evaluate students' use of the Mandala in studying a single play, it is possible to observe the following actions:

1. Were they able to load the play correctly?

2. Could they create the magnets they needed?

3. Having created the magnets, were they able to correctly explain what they were seeing?

4. Were they able to select items (speeches) for further study in the reading panel?

5. With their data in hand, could they provide an hypothesis that explained it?

6. Could they verify the hypothesis with further reading?

Features of Shakespeare's Tragedies

While the exercises using *Romeo and Juliet* (or any other play) allow the students to investigate questions that deal with a single play that they may have read in its entirety, this next exercise deals with studying a set of plays as a group. For some students these plays may all be familiar through close reading, but for others the visualizations in the Mandala might provide a first chance to consider all the plays together. The students might be familiar with one or more of the plays, or may be seeing all of them for the first time. In the first step, we load the Mandala with the eleven Shakespearean tragedies listed below. It is possible, of course, to also choose a subset of the tragedies based on some other criteria, or to add other plays as desired:

- *Romeo and Juliet*
- *Macbeth*
- *Hamlet*
- *Julius Caesar*
- *Othello*
- *Titus Andronicus*
- *Antony and Cleopatra*
- *Coriolanus*
- *Timon of Athens*
- *Troilus and Cressida*

There are three different ways to open all the plays at once in the Mandala: 1) load a single XML file that contains all the plays; 2) load a compressed archive (ZIP) file that contains the individual plays; or 3) merge each file one at a time into the Mandala. In all cases, it is interesting to see the variation in the number of speeches, which ranges from a low in *Titus Andronicus* (565) to a high in *Othello* (1181). The kinds of questions that can be addressed with a set of plays tend to begin with identifying trends. For instance:

- Are any of the acts noticeably shorter or longer than the others in terms of numbers of speeches?
- Are the key words that are related to central themes evenly distributed among the acts or among the plays?
- For those acts or plays with anomalous frequencies of occurrence of a keyword, what does further investigation of the speeches yield?

As an example, we will walk through an analysis of the linked concepts "perception," "cognition," and "expression" in the tragedies, using the frequently-occurring terms "see," "think," and "speak." To take advantage of the Mandala's ability to create regular expressions, we have also slightly expanded the search by adding a second term for each concept: "see or saw," "think or thought," and "speak or say". In Figure 9, the student has loaded all the speeches in all the tragedies. Note the density of dots around the periphery (10,456 in total).
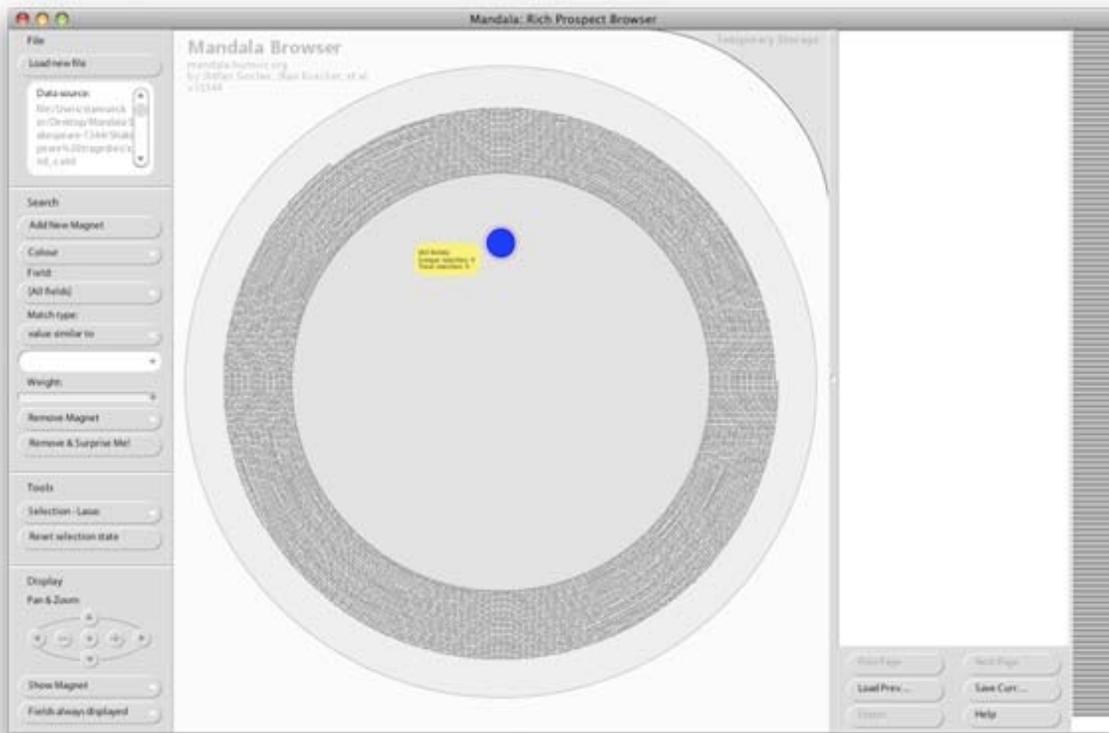
Figure 9: The 11 Shakespearean tragedies have been loaded as dots but only one blank magnet is visible.

Since we are interested in how these three concepts vary over the course of the tragedies, our next step is to create a magnet for each act (Figure 10). Everything appears to be working correctly in the XML tagging of the plays, since all the speeches are attracted from the periphery and there is no overlap between magnets. We can also see at a glance that Act 5 tends to be somewhat shorter in terms of numbers of speeches than the other acts (1801 speeches in Act 5 and over 2000 in each of the others). This suggests that Shakespeare's tragedies will tend to be perceived as moving somewhat swiftly to a conclusion, given the expectation for length that has been set up in the audience by the other acts.

Figure 10: The student has created one magnet for each of the five acts. These magnets can either be created one at a time, or else they can be created at one stroke by using field=Act and search term=[All Fields].

With this Mandala prepared, it is now possible to look at how "see or saw," "speak or say," and "think or thought" are distributed across the acts (Figure 11). The first thing we can notice is that "think or thought" occur in a total of 398 speeches, while "see or saw" are in 469 speeches, and "speak or say" are in a total of 836 speeches. It is therefore worth considering whether one of the features of the tragedies is that there is less explicit discussion of thinking going on among the characters than there is discussion of observation, and that there is more discussion about speaking than there is about either thinking or observing. To confirm these possibilities, it would be necessary to compare these results with a similar set of magnets for the romance plays or comedies or histories. It would also be worthwhile spending some time in looking for further synonyms, since it is possible that the key words we are using are not giving us the entire picture.
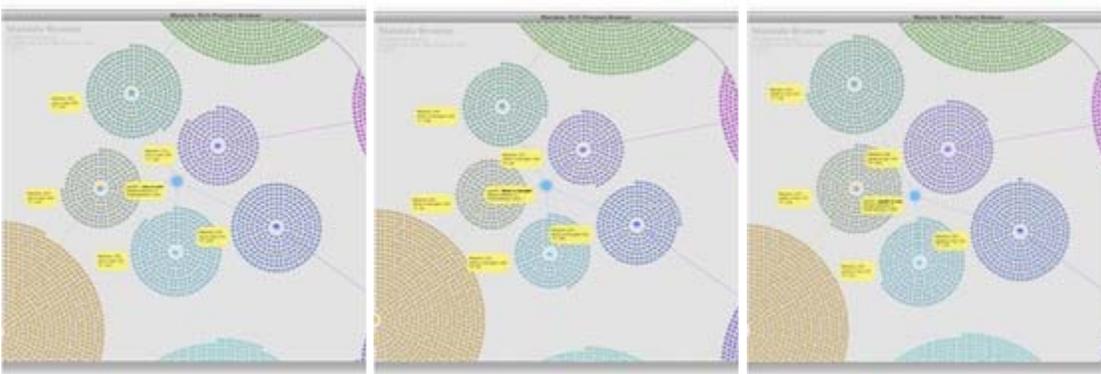

Figure 11: These three close-ups show the subsets, from left to right, of "see or saw," "think or thought," and "speak or say" in the five acts of the 11 tragedies.

Having spent some time considering the breakdown of the tragedies into acts, it is now worth turning our attention to comparisons among the plays. In Figure 12, the student has asked the Mandala to create a

magnet for each of the 11 plays. As we have previously noted, it is now possible to see the difference in the numbers of speeches in each play. *Othello* is the tragedy with the most total speeches (1181 in all), although *Antony and Cleopatra*, *Troilus and Cressida*, *Coriolanus*, and *Hamlet* are all close in size, with over 1100 speeches each. *Titus Andronicus* is the tragedy with the fewest speeches (total 565).
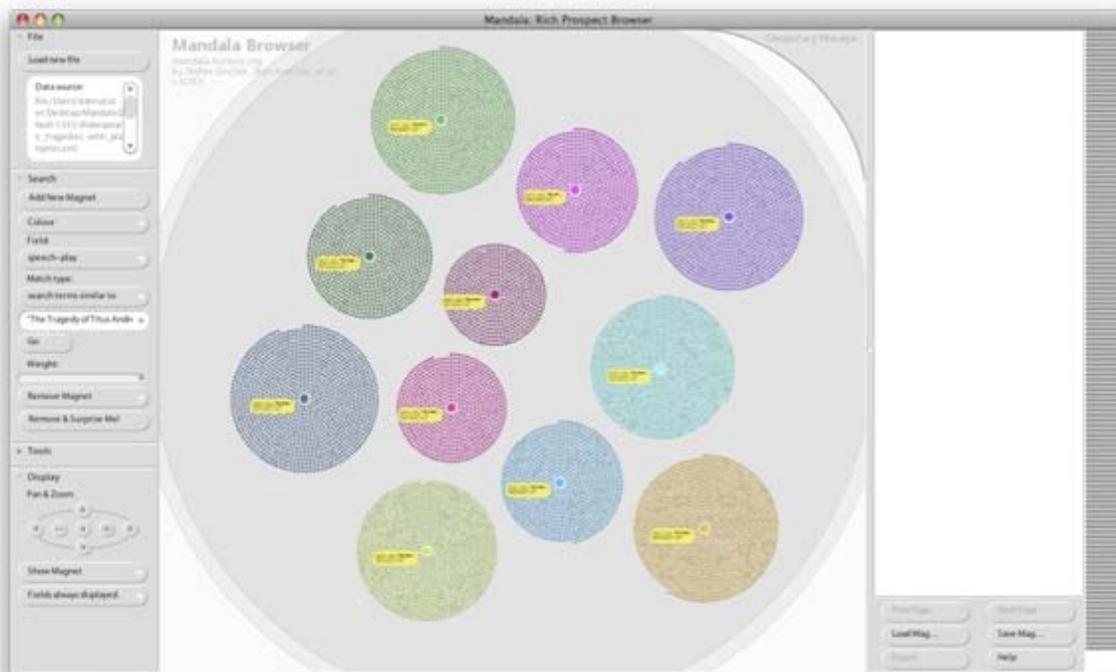

Figure 12: *In this screenshot, a magnet has been created for each of the 11 plays. Othello (left in blue) has the most speeches and Titus Andronicus (centre burgundy) the fewest.*

The next step (Figure 13) is to add a magnet for one of the key concepts – in this case, "speak or say". Of the 10,453 total speeches in all the plays, 836 contain at least one of these words. Since the plays differ significantly in numbers of speeches, it is important not to be misled by the size of the clusters around the pie magnets (which show the subsets). By adding the display of the ratios to the yellow labels, we can see the relevant percentages that can more properly be used for purposes of comparison. Of all the tragedies, *Titus Andronicus* has the most speeches (10%) that use the words "speak" or "say." *Macbeth* and *Coriolanus* are close seconds, with 9% each. Examining a few of the speeches at random suggests that "speak" and "say" tend to be imperative in general in the tragedies, with people either being ordered to speak or else not to speak. It might be worth further study, therefore, to see if these three plays are unusually focused on themes of expression or secrecy.
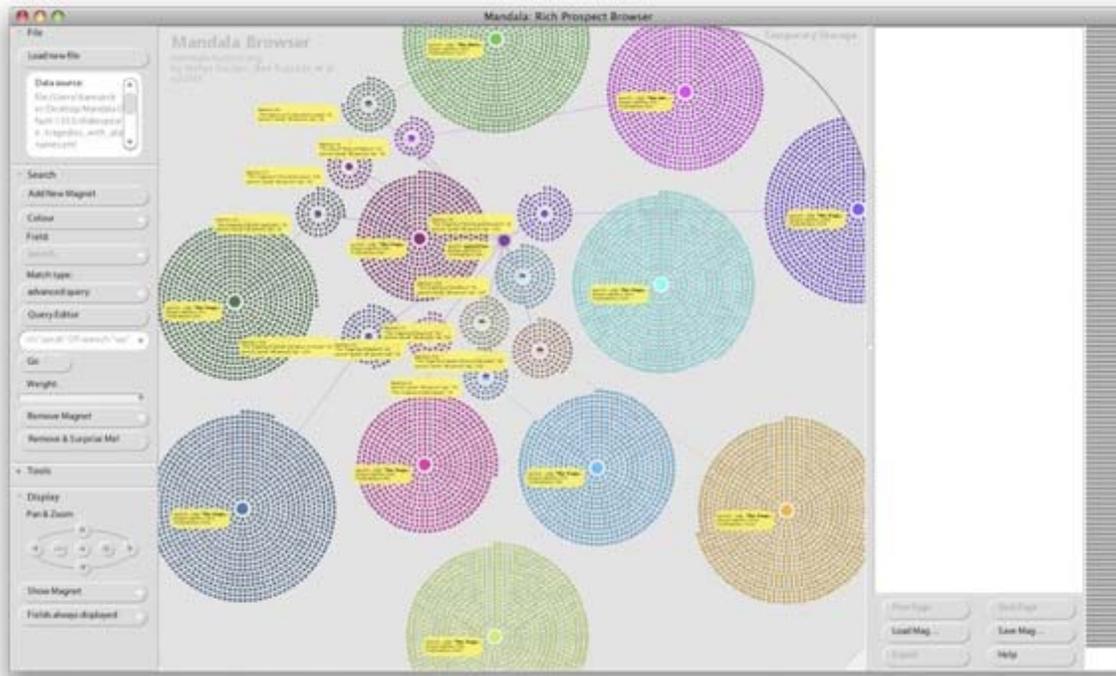
Figure 13: This Mandala shows the 11 plays and their use of the words "speak or say." The Mandala is beginning to get quite cluttered, but it is still possible to make some simple observations.

The next steps in the process would involve similar observations around the other keywords and their relative occurrence in each of the tragedies. At each point, it is important to follow up on the hypotheses being formulated by returning to the text of the plays and reading the speeches. In terms of evaluating students in their ability to carry out this kind of exercise, the possibilities are similar to those for using the Mandala with a single play.

Digital Reading

Although the Mandala Browser is a unique and powerful visualization tool for exploring digital texts, its functionality is focused on a relatively narrow band of search operations. The true potential of a digital text is that it lends itself to innumerable strategies for reading and analysis, since its constituent digital bits can be continuously rearranged and represented by computational tools. From very simple tasks (like a keyword search) to more complex tasks (like principal component analysis of semantic fields), the possibilities are only limited by the available texts and tools, and the imagination of their users. These tasks fall broadly under the rubric of text analysis, but the subset of tasks that we have presented using the Mandala Browser might be more usefully referred to as digital reading, oriented toward searching, browsing and reading text, without requiring more advanced knowledge of statistics and computational methods (see *A Companion to Digital Humanities* for several useful essays on better understanding text analysis for the humanities). What follows is a brief presentation of digital reading and suggested resources that would be complementary to using the Mandala Browser in the classroom (for a more in depth discussion of digital text tools in teaching, see Sinclair & Rockwell [2009]). Geoffrey Rockwell and Ian Lancashire (2005) provide an excellent overview of text analysis for the humanities (see also Rockwell [2003] for a more in-depth article on this topic): We can use computers to present, manage, and learn from electronic texts in ways difficult to do by hand. We can archive large quantities of text and make reliable copies of these archives. We can quickly retrieve passages from a large text database of millions of pages. We can ask where two or more words occur within the same paragraph. We can link automatically to other information from a hypertext. We can quantify writing style or try to identify the author of a disputed work

by his or her style. We can compare written works or study the evolution of language usage over a collection of texts. In general, the process of computer assisted text-analysis uses computers to search, retrieve, manipulate, measure and classify natural-language documents for patterns and by author, subject, and genre or type. Digital reading tools are concerned with providing incremental new functionality to conventional reading practices. Students and researchers in the humanities cannot be expected to fully abandon the familiarity of the sequential text -- nor can they be expected to embrace a different epistemological framework; digital reading tools are about enabling new modes of exploration and interpretation rather than, say, proving hypotheses about the formal characteristics of text. One of the most useful resources for digital reading is the list of recipies for exploration offered by the Text Analysis Portal for Research (TAPoR), which include a categorized list of instructions for various types of tasks, including the following:

- Identify simple themes within a text
- Explore colloquial word use in a text
- Analyze blog discourse

Many of the tools mentioned in the recipes are freely available online, including the Taporware tool suite (taporware.mcmaster.ca) and HyperPo (hyperpo.org). The main view of HyperPo, for instance, keeps the original digital text visible at all times and allows the user to interact with various data views, including word frequencies, concordances, and part of speech information (Sinclair 2003). Most online text analysis tools that allow users to provide their own text collections are designed to work with relatively small corpora up to a size equivalent to about three books. A notable exception to this is Voyeur Tools (voyeurtools.org), an online environment designed to scale to much larger corpora (see Figure 14). An innovative aspect of Voyeur Tools is that any of the results panels can be exported into remote web-based content and provide live functionality (this could be useful for students wishing to integrate tool results into blog posts, wikis or web-based essays (assuming the software supports the relevant markup tags).
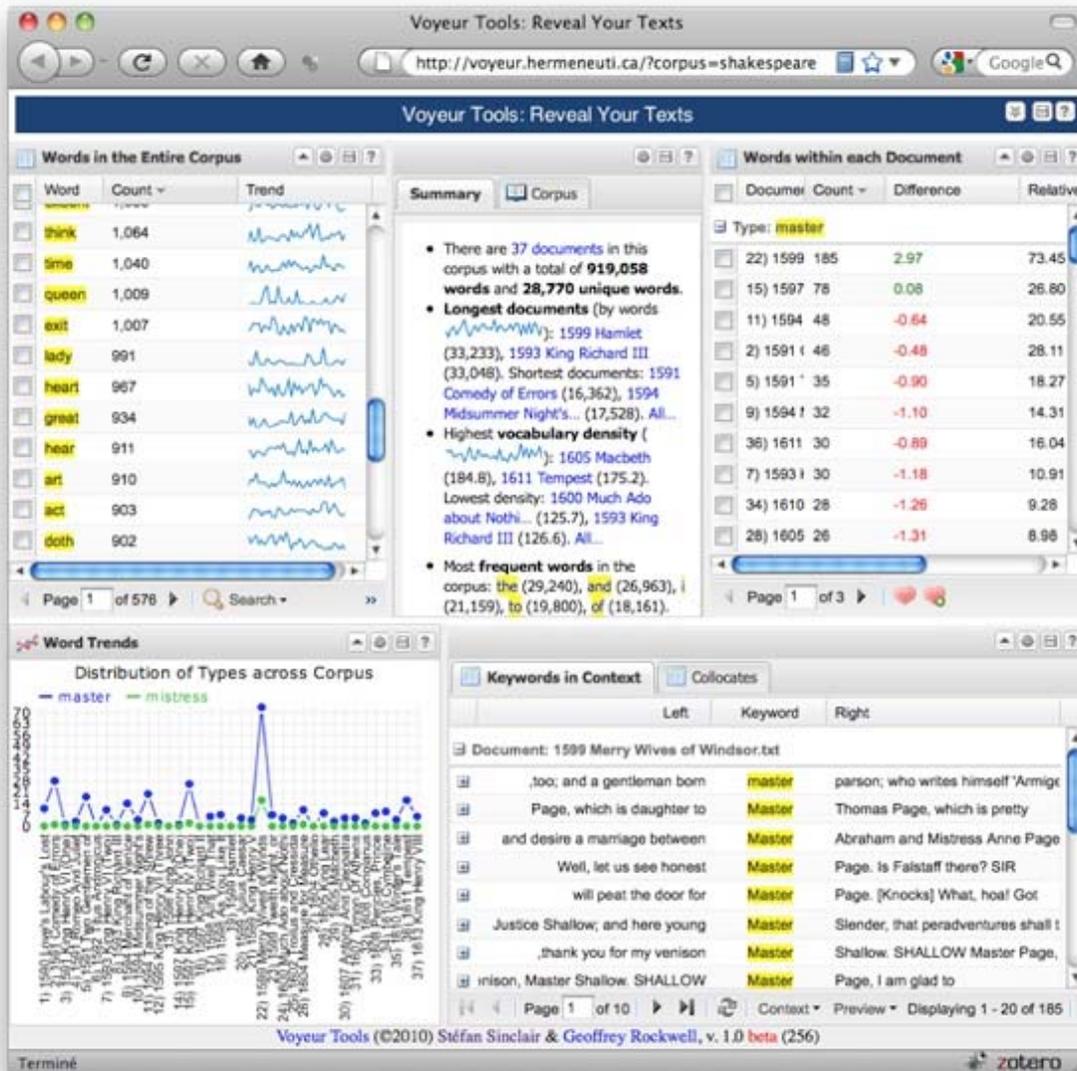
Figure 14: Voyeur Tools with 37 documents from Shakespeare (see address bar for URL).

Conclusions

In this paper, we have provided two examples of how students might work with the Mandala Browser to visualize XML-encoded versions of Shakespeare's plays. It is possible to develop some interesting hypotheses working with a single play, as in our first example, but it is in some respects more rewarding to deal with a set of several plays, which would be quite difficult to investigate without an interactive visualization. In that context, our second example deals with the subset of 11 of Shakespeare's plays that are commonly grouped together as the tragedies. The descriptions presented are meant only as examples of the kinds of interpretive processes enabled by the Mandala Browser; the texts offer endless further possibilities for exploration and analysis, especially when combined with other tools for digital reading, such as Voyeur.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *16*(7).

Bush, V. (1945). As we may think. *The Atlantic Monthly* , *176*(1), 101-108. Retrieved from http://www.theatlantic.com/doc/194507/bush.

Cheypesh, O., Pacher, C., Gabriele, S., Sinclair, S., Paulin, D., & Ruecker, S. (2006, May 29-31). Centering the mind and calming the heart: mandalas as interfaces. Paper presented at the Society for Digital Humanities (SDH/SEMI) conference, York University, Toronto.

Crane, G. (2006). What do you do with a million books? *D-Lib Magazine*, *12*(3). Retrieved from http://dlib.anu.edu.au/dlib/march06/crane/03crane.html

Gantz, J., et al. (2008). An updated forecast of worldwide information growth through 2011. *IDC*. Retrieved from http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-....

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*.

Manovich, L. (2009, June 22-25). Cultural analytics. Plenary address presented at the Digital Humanities conference, University of Maryland.

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history. London: Verso.*

Ramsay, S. (2003). Toward an algorithmic criticism. *Literary and Linguistic Computing*, *18*(2).

Rockwell, G. (2003). What is text analysis, really? *Literary and Linguistic Computing*, *18*(2).

Rockwell, G., & Lancashire, I. (2005). What is text analysis? Retrieved from http://tada.mcmaster.ca/Main/WhatTA.

Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004). *A Companion to Digital Humanities*. Retrieved from http://www.digitalhumanities.org/companion/. Oxford: Blackwell.

Sinclair, S. (2003). Computer-assisted reading: Reconceiving text analysis. *Literary and Linguistic Computing*, *18*(2).

Sinclair, S., & Rockwell, G. (2009). Between language and literature: Digital text exploration. In I. Lancashire (Ed.), *Teaching Literature and Language Online*. New York: MLA.

Shakespeare, W. *The Complete Plays*. The Nameless Shakespeare collection at WordHoard, Northwestern University. Retrieved from http://wordhoard.northwestern.edu/.

Unsworth, J. (2005). New methods for humanities research. *Lyman Award Lecture* Retrieved from http://www3.isrl.uiuc.edu/~unsworth/lyman.htm.