DHQ: Digital Humanities Quarterly
2009
Volume 3 Number 3

Designing Data Mining Droplets: New Interface Objects for the Humanities Scholar

Stan Ruecker <sruecker_at_ualberta_dot_ca>, University of Alberta, Canada
Milena Radzikowska <mradzikowska_at_gmail_dot_com>, Mount Royal College, Canada
Stéfan Sinclair <sgs_at_mcmaster_dot_edu>, McMaster University, Canada

Abstract

In this paper, we describe the design of a number of alternative interface “droplets” that are intended for use by humanities scholars interested in applying data mining and information visualization tools to the task of hypothesis formulation. The trained droplets provide several functions. Their primary purpose is to encapsulate the results of the software training phase. They can be saved for future re-use against other collections or combinations of collections. They can be modified by having the user accept or reject features identified by the data mining software. Finally, they can also contain choices for how to display and organize items in the collection. The opportunity to develop a new interface object presents the designer with the challenge of effectively communicating what the tool is good for and how it is used. This paper outlines the design process we followed in creating the visual representations of these interface objects, describes the communicative strengths and weaknesses of a number of alternative designs, and discusses the importance of the study of new interface objects as the means of providing the user with new interface affordances.

Introduction

The goal of this paper is to address some of the conceptual issues that arise in the design of a new kind of interface object for a specific domain — data mining for the humanities. In that context, we describe one component of our research: the design of a form of visual representation that would provide humanities scholars with some insight into the data mining process, while at the same time making the activity of data mining attractive and easy to carry out.

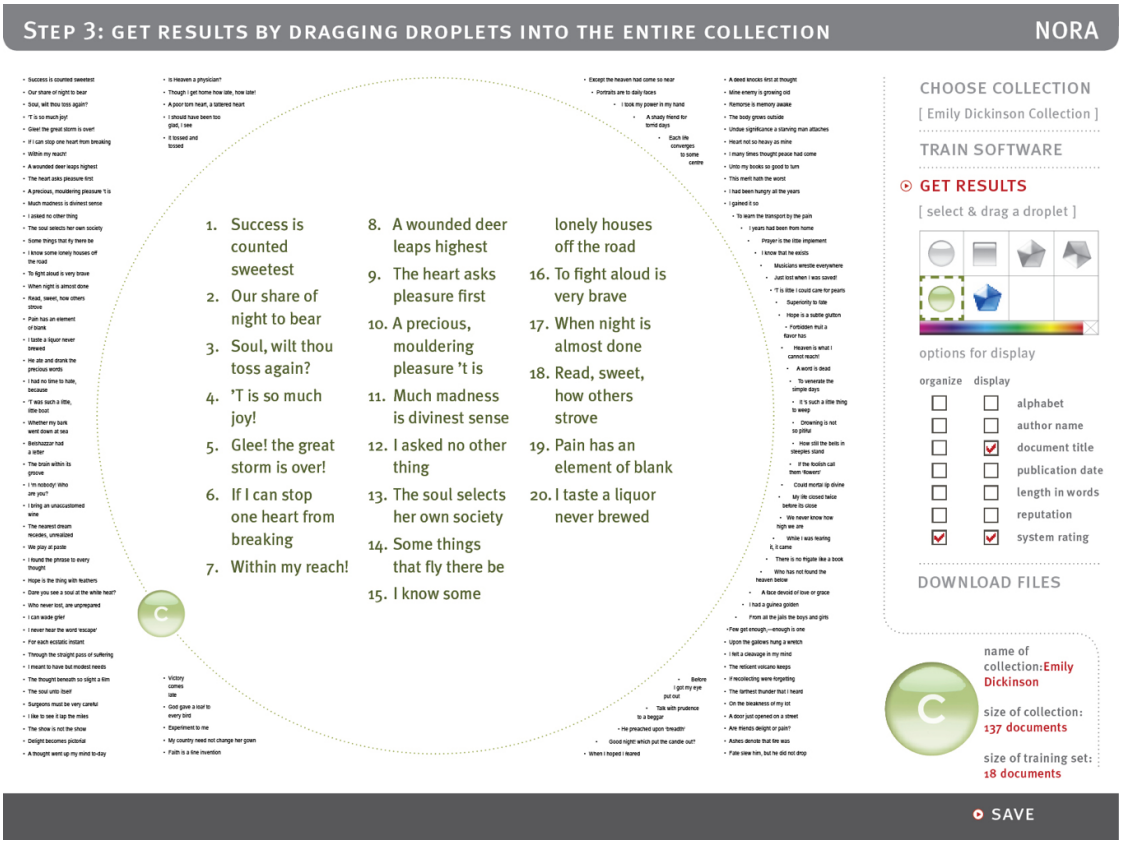

Figure 1. An early sketch of the data mining environment (in this case for the NORA project) shows someone using a droplet trained for identifying the erotic in a set of poems from the Emily Dickinson collection in the Institute for Advanced Technology in the Humanities (IATH) at the University of Virginia. Note that the preliminary droplet design shown here (bottom right) has no specifically communicative morphology.

Our strategy in this interface was to provide the user with a variety of empty “droplets” which would be filled with the results of the software training phase [Ruecker et al.]. Each droplet would contain or encapsulate an entire working state of the system, including the algorithmic consequences of a particular training exercise, combined with some parameters for organizing and selecting the form of the display. The choice of the proper word to identify the droplets is in itself a subject of design. Other terms that have been suggested include “magnet,” “crystal,” “capsule,” “lens,” “charm,” “filter,” “system state,” “kernel,” and the very Canadian “hockey puck.” Whatever these objects are eventually called, for the time being we are using the term “droplet,” which suggests to us a densely compressed item that can unpack in an organic way to influence the entire surroundings. Once a droplet has been trained for data mining, it can be saved and applied to the entire collection, or to a different collection. A droplet is applied to a collection by dragging and dropping it onto a display representing each item, after which the display organizes itself in a series of “oil and water” effects. 2

Of vital significance to the success of this strategy is the design of the droplets. The droplets need to be able to represent the relevant information about the data mining process in a form that is readily interpretable by humanities scholars. The droplet serves in one sense like an icon — a person looking at it will hopefully remember what system state it contains. This iconic function should work at different scales, at least one of which is quite small. The droplets therefore need to be easily visually differentiable one from another, at every scale. Finally, the droplets need to be visually appealing. We describe here our initial attempts to design these interface objects, based on a set of metaphors to real-world items that combine complex visual appearance with a compact form. 3

Background

The online availability of a wide range of digital data has resulted in a corresponding increase in various kinds of tools for retrieving and manipulating the items in a collection [Hockey 2000]. Interface design researchers have worked on systems intended to help users access digital images, work with electronic text files, and apply data mining algorithms 4

to a variety of problems, both in the sciences and in the humanities.

In the area of digital images, [Bederson 2001] describes a zoomable browser, [Bumgardner et al. 2005] provides an experimental search tool that uses a colour wheel as its interface, and [Hascoët et al. 1998] discuss the use of maps in accessing a digital library. Other examples include [Rodden et al. 2001], who studied the use of similarity clustering for browsing tasks, and [Ruecker et al. 2005], who developed a prototype image browser for pill identification. 5

For tools related to text files, [Pirolli et al. 1996] describe a system for visualizing documents which allows the user to form dynamic groups. [Small 1996] developed a 3D prototype for text navigation, where the reader moved between columns of text from Shakespeare's plays. A variety of researchers have worked in the area of data mining for text collections of various kinds. For example, [Feldman et al. 1997] discuss early efforts in this area, and [Weiss et al. 2005] provide a recent update on methods. 6

Some researchers have pointed out that the potential for applying data mining tools to questions in the humanities lies largely in the capacity of such tools to contribute, not primarily to hypothesis testing, but instead to hypothesis formulation [Shneiderman 2001]; [Ramsay 2003]; [Unsworth 2004]. The standard approach in humanities research is not to solve a problem by testing one hypothesis against another, but rather to enrich the object of study by repeated observation and reporting. Data mining tools and their accompanying visualizations, which facilitate pattern finding across a wide range of data, can definitely play a role in this process. 7

With respect to the design of interfaces for data mining, it is important to remember that each new online tool represents a new opportunity for action, or affordance [Gibson 1979]; [Vicente 2002]. For instance, in a more conventional approach to the interface for data mining, it would be possible to create a history palette that records previous states of the system. However, it is not necessarily straightforward to repurpose an item from that history to a new collection. By encapsulating the history states as droplets, we make the repurposing simpler. 8

Another significant feature of the droplets is their role in interactivity. By providing the user with an item to drag and drop to trigger a series of dynamic responses from the system, the droplets help facilitate an instructional aspect: the user can see the steps carried out by system, which correspond to the steps associated with the droplet. While visually dynamic responses are not reliant on the presence of droplets as objects, their existence as part of the user interaction helps to suggest to the designer these various new forms of feedback, which are a kind of affordance. 9

Studying these new affordances presents a challenge, in that the researcher by definition does not always have an existing object with a similar affordance — otherwise it would be a case of a redesign rather than a new tool [Ruecker 2003]. Though opinions vary, the current dominant perspective is that interface research requires a component of usability study [Nielsen 2000], but that usability study alone is probably not enough. Attention should also be paid to other factors, such as aesthetics [Karvonen 2000], effect [Dillon 2001], and sustained use over time [Plaisant 2004]. 10

Methodology

We began by identifying the kinds of information the user might want to know while working with the system. These included an overview of the process, suggestions about the kinds of tasks that could be performed using the system, reassurance at each point that the right things were happening, and assistance in interpreting the results of each stage and moving successfully to the next stage. With the droplets, we hoped to be able to communicate what had been done to create them, in order to suggest how they might be successfully deployed once they were created. 11

To construct the droplets, we generated a candidate list of real-world items that have a sufficiently complex physical shape to serve as possible metaphors for the complexities of the data mining process. We determined early in the process that it would be difficult and probably not helpful to attempt to communicate for this demographic the actual algorithms involved, as for example by superimposing an equation on a geometric shape. Instead, we hoped to be able to visually express the following information: 12

- Is this a trained droplet or an empty one?
- For trained droplets, has the user accepted the features recommended by the system or has the list of

features been modified?
- What kinds of features were included?
- How many features were included?
- What options for display have been associated with the droplet?
- What choices for organizing the display have been applied?

There are also other pieces of information that could be useful for understanding what has been happening. These items need to be communicated somehow but could be difficult to associate with the visual appearance of the droplets. These include: 13

- The name of the collection or collections used in training.
- The size of the collection.
- The size of the training set.
- The name and goals of the person responsible for training the droplet.

Some strategies involving droplet morphology might include using the size of the droplet to indicate the size of the training set or of the collection the set was drawn from. Internal and external lines can also be thickened or lightened as a way of suggesting robustness of the training set. Finally, depending on the visual kind of droplet, it may be possible to nest one droplet inside another, as a way of indicating their use in combination. 14

It may also be possible to associate this information with the droplets using strategies that do not involve the droplet morphology *per se*, but instead rely on the combination of text and image. Combining these methods is seen by some theorists as an important approach to the design of technical communications [Horn 1998]. We will provide this connection in the case of the prototype by refreshing an information panel about the droplet details whenever the user selects a droplet. This panel will also provide the opportunity to adjust some of the settings stored by the droplet. 15

Results

Working from our original map of over a dozen potential metaphors (Figure 2), we selected the following short list for further investigation. We wanted to have a variety of items that were distinct from each other but were also visually complex in a way that could communicate the stages in droplet training. We thought we should include examples that covered points on a terrain that included the organic and the mechanical, with reference to several disciplines. Finally, we tried to choose examples that could be contained by a common perimeter. Our working list contained the following items: 16

- Ferns — configurations of individual organic pieces that form larger items
- Snowflakes — a single solid unique configuration that relies on symmetry
- Solar system — individual items in relations suggested by a larger structure
- Atoms — individual items connected in a more elaborate geometric framework
- Cells — complex interiors composed of pieces that associate by juxtaposition
- Clockwork — complex interiors consisting of structures that interconnect
- Lego™ — geometric shapes with complex surfaces that interconnect

For each of these metaphors, we developed sketches for four different states of the droplet: untrained, trained, trained with multiple display options chosen, and trained with multiple display and two different organization options. Our goal in each case was to make the different states visually distinct at every level of magnification, and to make the number of display and organization options obvious at the largest size. 17

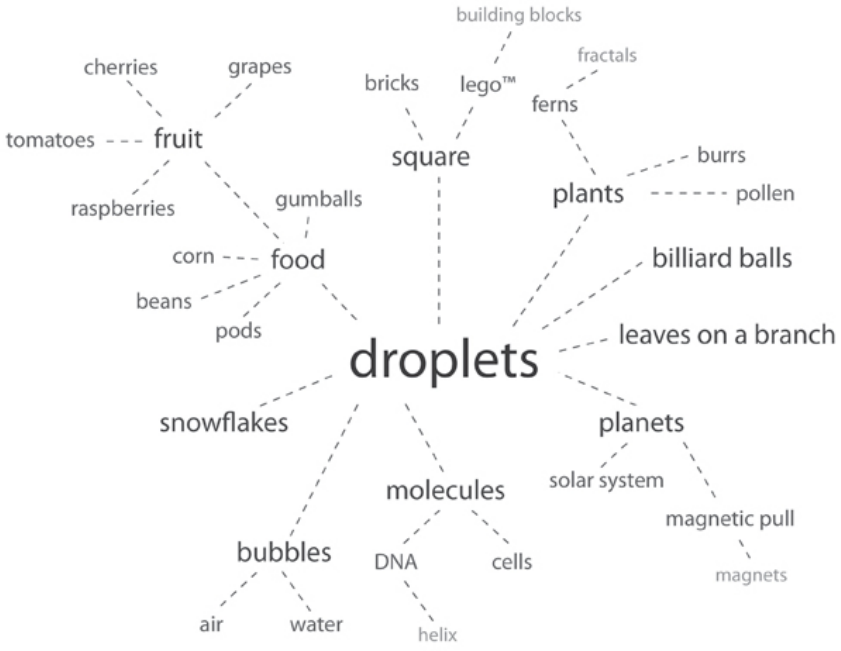

Figure 2. Our concept map of possible droplet metaphors shows a wide range of candidate real-world objects that combine visual complexity with a compact form.

We chose these various states because they represent significant choices made by the user. It would also be possible to consider visually representing choices the user makes about what collection to work with in the first place, which may be one of the most significant choices the user makes. However, visually representing collections is definitely a challenge, and it may be preferable to provide information about the collection in the form of text labels. 18

Ferns

A fern is a fractal, which means it repeats its morphology at increasing scales (Figure 3). We might adopt this strategy for two scales, where in the unfolding fern leaf, the individual leaflets represent functions and the entire leaf represents the complete, organized droplet. 19

We can use the stem to represent the software training, and the leaflets to represent the other functions. This strategy has the benefit of looking minimal when no display or organization functions are chosen, which may prompt the user to want to choose more sophisticated configurations of options. 20

If we also assume that the two sides of the stem represent two kinds of organization, then having all the display items on one side of the stem would indicate only one kind of sorting, while dividing display items on both sides of the stem would indicate two kinds of sorting. 21

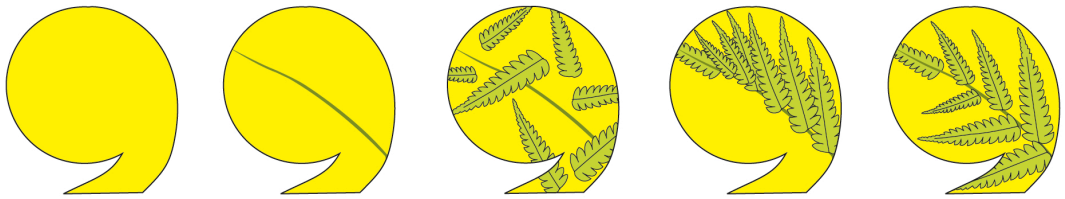

Figure 3. The placement of leaflets along the stem of the fern leaf allows us to express the user choices starting with an empty droplet (left), then sequentially adding training data, display choices, then organization in one way and in two ways.

Reading the sequence from left to right, we show first an untrained or empty droplet. The next version shows a droplet that has been trained by the user. Taking one of the demonstration projects as an example, this second droplet might contain the results of training the system to recognize poems by Emily Dickinson with an erotic charge, using a naïve Bayesian algorithm. The third version shows this same trained droplet with seven items chosen for display. In the case of the Dickinson collection, these items might include the poem's title (often the first line), the date of first publication, the place of publication, the name of the publisher, the number of lines in the poem, the number of words in the poem, the number of key features found in the poem related to eroticism, and the numeric score assigned by the system for the poem in terms of its erotic charge. The fourth version would represent the same information about each poem, but organize the results in some way — perhaps by the numeric rating assigned by the system. The fifth and final version would show the items arranged in two ways — first by numeric rating, and chronologically within that. 22

The organic nature of the fern droplet may lead to some difficulties for the user in that a growth process for a fern is not the same as selection among various options by a user defining a droplet. The use of this organic metaphor, however, does suggest another possibility — would it be interesting to indicate how long it has been since someone used a droplet? Do the droplets visibly age when they aren't used? Does new use refresh the appearance of the droplet? Would people be encouraged to experiment with strange droplets because they are obviously drying up or deteriorating? 23

Snowflakes

Ferns suggest quite a regular form of arrangement, which means there is little meaningful variation possible between different droplets. Snowflakes also tend to symmetry, but each is unique. They combine a complex silhouette with a compact form (Figure 4). Variations in the details comprising the silhouette could therefore be used to communicate a wide range of functions. 24

However, the strong visual language of the snowflake may prove to be difficult to repurpose as a meaningful channel of communication. The fact that each snowflake is supposed to be unique also means that there is no basic, restricted vocabulary of shapes to draw on in their construction. 25

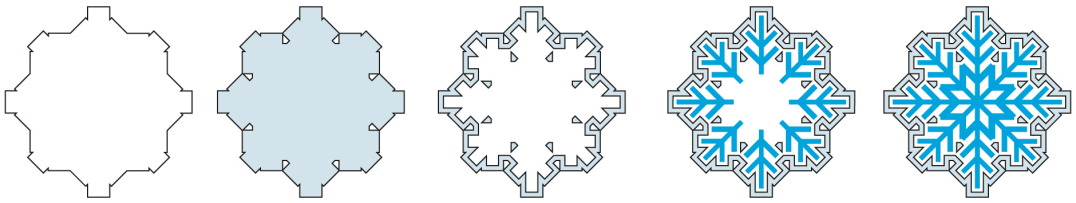

Figure 4. Each snowflake is a unique visual object, which allows us to differentiate one droplet from another, but introduces a difficulty in that there is no simple method of re-using recognizable components.

Our draft solution in this case is to treat the visual complexity of the interior of the object as the measure of the state of 26

the droplet. Unlike our other designs, which involve composites of countable objects, the snowflake droplets indicate each condition by filling in spaces that are otherwise unarticulated.

Solar System

Objects in the solar system create a composite object where the individual items are in relation to one another but not in immediate contact (Figure 5). The central position of the sun also serves to imply the centrality of the software training. A solar system without a sun is clearly incomplete. 27

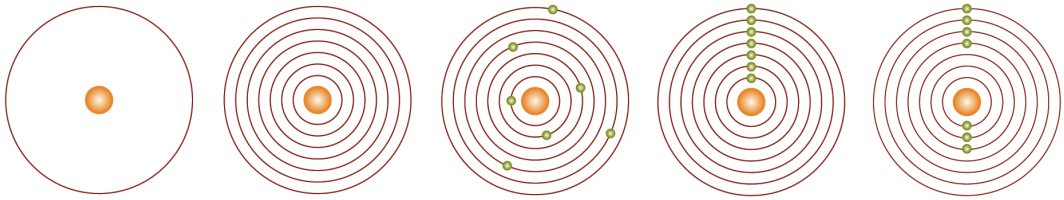

Figure 5. The solar system, with its objects in orbit, provides a structure that can be progressively filled with planetary dots that represent choices of representation, while location on the orbits is used to indicate organization.

Another potential difficulty with several of the designs, including the solar system, is that they may suggest a degree of order and regularity which may be somewhat at odds with the experience of the scholar using data mining techniques. Using a data mining system can actually involve an iterative and somewhat “messy” experimentation with various options. 28

Atomic

Our starting point for the atomic droplets are the simple models that consist of electrons in elliptical orbits around a nucleus (Figure 6). The nucleus is filled in during the training phase, while the inclusion of electrons and their locations represent choices about item representation and organization. 29

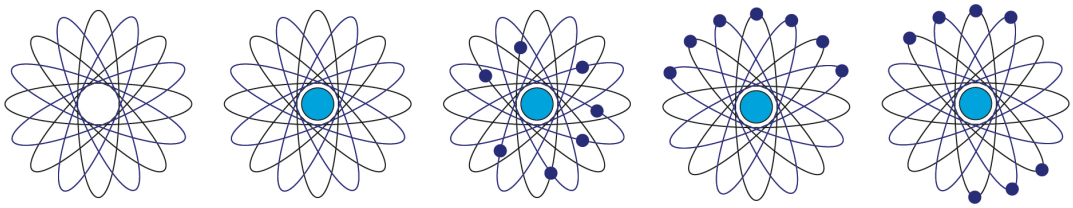

Figure 6. Atomic models provide a vocabulary for expressing the components of the droplets, consisting of individual items connected to each other.

Cells

A cell has an interior that is populated with a number of distinct individual items and structures (Figure 7). Cells therefore provide a compact metaphor based on the complexities of the interior of the droplet. We also have available for future exploration the single-celled organisms, such as the paramecium, which combine this interior complexity with an exterior with some communicative potential. 30

Cells also suggest an organic form, which may help to counterbalance the highly technical profile of data mining in the humanities. 31

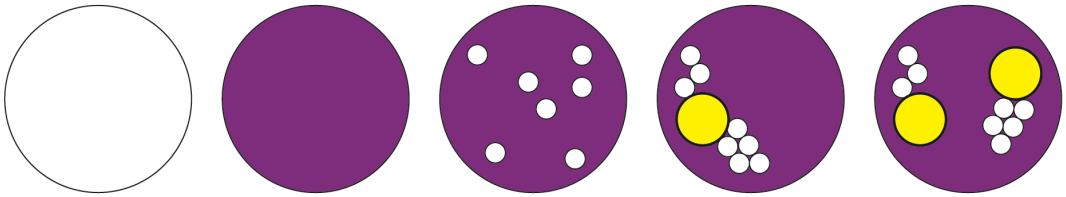

Figure 7. A cell is neither an aggregate nor does it have a complex silhouette. Its communicative potential consists instead of a rich interior of organic shapes, including individual items and structures that divide, enclose, and support them.

Clockwork

A clockwork is a complex interior like a cell, without the suggestion of the organic (Figure 8). There is a high degree of interconnection of the parts inside a clock, implying that all the parts are necessary in order for it to work. This level of constraint on what is necessary and what is optional might not be appropriate in the context of data mining, but the operational nature of the clock and the implied association with the mathematical operations underlying data mining may make it particularly appropriate. 32

The variety of interior components also provides a potentially rich visual vocabulary for representing the different aspects of the droplets. Finally, we have used an external outline suggestive of clock gears, in order to allow a direct visual association to the mechanical, even for the untrained form of the droplet. 33

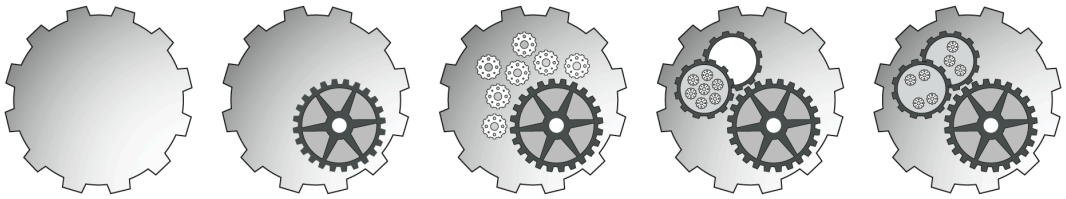

Figure 8. Like a cell, a clockwork shows a rich internal landscape that can be used to represent a variety of functions. Clockworks are mechanical rather than organic, and therefore suggest interconnection, rather than isolation of the functions.

Lego™

With Lego, there are a set number of individual shapes that are aggregated. With this metaphor, we can use the external contour of the composite droplet (Figure 9). We can distinguish by size between more and less important functions, so the central training can be indicated by large Lego piece, while the display functions are secondary and the organization functions tertiary. 34

Lego also comes with the affordance of assembling the separate pieces into different configurations. The user could distinguish between similar droplets by taking advantage of different kinds of arrangement. 35

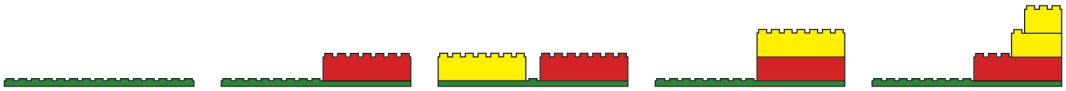

Figure 9. Lego™ suggests a method of combining separate items to create a new whole. For our purposes, each individual piece of Lego would stand either for the result of software training or for a choice of representation or organization.

Conclusions and Future Research

Having identified a range of possibilities, our next step will be to present them to potential users in order to collect measures of performance and preference. By placing them in the interactive context of a prototype environment, we will be able to examine how humanities scholars respond to the various affordances. The goals of this phase will be to determine whether participants are able to make the necessary intuitive leaps to understand the intended communicative aspects of each of the droplet designs. Once we've established a smaller subset of droplets, we will proceed by expanding the visual positioning or skinning of each droplet type, in order to determine how humanities scholars respond to various semantic differentials such as glossy/rough, technological/natural, geometric/organic, and colour/grey scale. By determining how potential users of the data mining system perceive the design dimensions of the droplets, we will be able to decide to what extent this strategy can prove beneficial in removing barriers to them adopting the system. One possibility may consist of the use of a hybrid form of droplets, where different visual components are assembled in a kind of toolkit. Our eventual decisions with respect to the design of the droplets may also be usefully repurposed to inform the visual aspects of the design of the entire system. 36

Acknowledgements

The authors wish to thanks the many members of the NORA project research team for their contributions to this work. Their names can be found at http://www.noraproject.org/team.php. We would also like to acknowledge the generous support of the Andrew W. Mellon Foundation, the Social Sciences and Humanities Research Council of Canada, the Natural Sciences and Engineering Council of Canada, and the Canadian Foundation for Innovation. 37

Works Cited

Bederson 2001 Bederson, B.B. "PhotoMesa: a Zoomable Image Browser Using Quantum Treemaps and Bubblemaps". Presented at *ACM 2001. Proceedings of the 14th Annual ACM Symposium on User Interface Software Technology* (2001), pp. 71-80.

Bumgardner et al. 2005 Bumgardner, J. *Flickr Color Fields Experimental Colr Pickr*. 2005. http://krazydad.com/colrpickr/.

Dillon 2001 Dillon, Andrew. "Beyond Usability: Process, Outcome and Affect in Human-Computer Interactions". *Canadian Journal of Library and Information Science* 26: 4 (2001), pp. 57-69.

Feldman et al. 1997 Feldman, Ronen, and Haym Hirsh. "Finding Associations in Collections of Text". In Ryszard S. Michalski Ivan Bratko and Miroslav Kubat, *Machine Learning and Data Mining: Methods and Applications*. New York: Wiley, 1997. pp. 223-240.

Gibson 1979 Gibson, James J. *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin, 1979.

Hascoët et al. 1998 Hascoët, Mountaz, and Xavier Soinard. "Using Maps as a User Interface to a Digital Library". Presented at *SIGIR '98. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1998), pp. 339-340. http://doi.acm.org/10.1145/290941.291028.

Hockey 2000 Hockey, Susan. *Electronic Texts in the Humanities*. Oxford: Oxford University Press, 2000.

Horn 1998 Horn, Robert E. *Visual Language: Global Communication for the 21st Century*. Bainbridge Island, WA:

MacroVU Inc, 1998.

Horton et al. 2006 Horton, Tom, Kristen Taylor, Bei Yu and Xin Xiang. “Quite Right, Dear and Interesting: Seeking the Sentimental n Nineteenth Century American Fiction”. Presented at *Digital Humanities 2006*. *Proceedings of the Association for Literary and Linguistic Computing* (2006), pp. 81-82.

Karvonen 2000 Karvonen, Kristiina. “The Beauty of Simplicity”. Presented at *CUU*. *Proceedings of the 2000 Conference on Universal Usability* (2000).

Kirschenbaum et al. 2006 Kirschenbaum, Matthew G., Catherine Plaisant, Martha Nell Smith, Loretta Auvil, James Rose, Bei Yu and Tanya Clement. “ ‘Undiscovered Public Knowledge’”: Mining for Patterns of Erotic Language in Emily Dickinson's Correspondence with Susan Huntington (Gilbert) Dickinson”. Presented at *Digital Humanities 2006* (July 5–9, 2006). *Digital Humanities 2006 Conference Abstracts*, pp. 252-255.

Nielsen 2000 J. Nielsen. *Designing web usability: The practice of simplicity. Indianapolis*, IN: New Riders, 2000.

Pirolli et al. 1996 Pirolli, Peter, Patricia Schank, Marti Hearst and Christine Diehl. “Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection”. Presented at *SIGCHI 2006*. *Proceedings of the SIGCHI conference on Human factors in Computing Systems: Common Ground* (1996), pp. 213-220.

Plaisant 2004 Plaisant, Catherine. “The Challenge of Information Visualization Evaluation”. *IEEE Proceedings of AVI 2004* (2004).

Ramsay 2003 Ramsay, Stephen. “Toward an Algorithmic Criticism”. *Literary and Linguistic Computing* 18: 2 (2003), pp. 167-174.

Ramsay and Steger 2006 Ramsay, Stephen, and Sara Steger. “Distinguished Speakers: Keyword Extraction and Critical Analysis with Virginia Woolf's The Waves”. Presented at *Digital Humanities 2006*. *Proceedings of the Association for Literary and Linguistic Computing Conference 2006* (2006), pp. 255-257.

Rodden et al. 2001 Rodden, Kerry, Wojciech Basalaj, David Sinclair and Kenneth Wood. “Does Organization by Similarity Assist Image Browsing”. Presented at *CHI 2001*. *Proceedings of the Human Factors in Computing Systems Conference* (2001), pp. 190-197.

Ruecker 2003 Ruecker, Stan. *Affordances of prospect for academic users of interpretively-tagged text collections*. Thesis, University of Alberta, Edmonton, Alberta, Canada: 2003.

Ruecker et al. Ruecker, Stan, Milena Radzikowksa and Stéfan Sinclair. “Communicating Process with Form: Designing the Visual Morphology of the Nora Data Mining Kernels”. Presented at *CaSTA 2006*. *Proceedings of the Joint Computer Science and Humanities Computing Conference* (2006), pp. 57-68.

Ruecker et al. 2005 S. Ruecker, L. M. Given, B. Sadler, and A. Ruskin. “Building Accessible Web Interfaces for Seniors: Similarity Clustering of Pill Images.” *Include 2005. London. Helen Hamlyn Institute. Royal College of Art.* April 5-8, 2005, 2005.

Shneiderman 2001 Shneiderman, Ben. “Inventing Discovery Tools: Combining Information Visualization with Data Mining”. Presented at *DC 2001*. *Keynote for Discovery Science Conference 2001* (2001).

Small 1996 Small, David. “Navigating Large Bodies of Text”. *IBM Systems Journal* 35: 3-4 (1996).

Unsworth 2004 Unsworth, John. “Forms of Attention: Digital Humanities Beyond Representation”. Presented at *CaSTA 2004*. *Proceedings of the Third Conference of the Canadian Symposium on Text Analysis* (2004).

Unsworth 2005 Unsworth, John. “New Methods for Humanities Research”. Presented at *National Humanities Center*. *The 2005 Lyman Award Lecture* (2005). http://www3.isrl.uiuc.edu/~unsworth/lyman.htm.

Vicente 2002 Vicente, Kim J. “Ecological Interface Design: Progress and Challenges”. *Human Factors* 44: 1 (2002), pp. 62-78.

Weiss et al. 2005 Weiss, Sholom M., Nitin Indurkhya, Tong Zhang and Fred Damerau. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer, 2005.